Neural Network Control for Active Cameras Using Master-Slave Setup

Renan Oliveira Reis, Igor Dias, William Robson Schwartz
Smart Surveillance Interest Group, Smart Sense Laboratory, Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
renanreis1@ufmg.br, igorhenrique94@ufmg.br, william@dcc.ufmg.br

Abstract

The use of active cameras has increased to perform tasks such as tracking and biometrics at distance. Furthermore, recent efforts have focused on the master-slave setup, which is composed of fixed and PTZ cameras. Although, there are many works regarding active camera control, there is no standard way to compare different control approaches once the experiment cannot be reproduced. Thus, in this work, besides the proposition of a novel learning-based approach to the master-slave setup, we also propose an experimental setup that allows a fair comparison between different methods. The proposed control method learns corresponding points between the fixed and the PTZ cameras using a neural network. The novel experimental setup places two PTZ cameras side-by-side with a very similar view so that two different algorithms can be executed simultaneously. The experiments show that the proposed method is better than literature method when the focus is centralizing a target at the PTZ view.

1. Introduction

Initially, simple surveillance systems were composed of a single fixed camera placed at strategic locations of surveillance space. However a limited number of camera presents problems as complete coverage and generation of low resolution images when the targets are distant from the cameras. To solve these problems is possible to use a network connected with multiple fixed cameras so more areas and target can be observed with more detail. Although multiple fixed camera solve the problems of limited surveillance space, the costs with processing, hardware and devices are high and the problem of low resolution of target can not be solved. Therefore, a possible solution of this problem is employing PTZ cameras (active cameras), which can adjust its view by varying the Pan, Tilt, and Zoom (PTZ) parameters according to the target location. The combination of fixed and active cameras provides an hybrid setup that might allow a larger coverage of the scene at reduced costs and can provides high-resolution images of targets at distance [10, 1].

In the setup composed by fixed and PTZ cameras, called master-slave, the fixed camera has a wide view of surveillance space and a PTZ camera acts as a slave by executing commands received according to the target's location. The correspondence between master and slave camera is a process that can be generated by a mapping between these two cameras. Some strategies to handle the correspondence such as calibration and setup manual become the process onerous and demands efforts because requires cooperation between person and cameras or introduction of marks or patterns in the surveillance space. In addition, the calibration needs to be repeated whether the camera setup changes due external factors such as maintenance or vandalism. Thus, a smart way to create a correspondence in master-slave setup is necessary to improve the robustness for this type of systems [10, 5, 14, 11].

Due the master-slave correspondence is possible to observe target with details at distance from the camera. However, it is difficult to measure the methods progress. Reproducing experiments in master-slave setup requires more efforts because online tests limit the possibility of repetition and there is no standard protocol or methodology. The current literature contributes with virtual environments or virtual cameras concentrated in single active camera [5, 8, 13], which shows an extend area to explore.

In this paper, we propose a novel learning-based method to control master-slave setup in surveillance scenarios. We introduce an automatic algorithm capable of recording information from targets moving in the scene, and then, a designed neural network learns a model that represents the correspondence between master and slave cameras. Afterwards, we employ the learned model in a real and outdoor scenario to provide tracking with high-resolution images anywhere in the covered scene.

To handle the performance evaluation problem of master-slave control approaches, we propose a novel experimental setup where two PTZ cameras are placed side-by-side with a similar view, allowing the execution of two different algorithms simultaneously in the same scene, pro-
providing a fair comparison. To the best of our knowledge, this is the first attempt to compare two control methods running simultaneously in the exactly scenario and in real-time. We compare our approach with the control method proposed by Wang et al. [14] and the results show that our method achieves better performance.

2. Related Work

Recent studies have applied efforts in automatic video surveillance acquiring high resolution images of targets in the scene using active cameras combined or not with fixed cameras [14, 3, 4, 11, 7]. In contrast, less works have focused on efficient evaluation metrics and protocols [5, 8, 13].

Wang et al. [14] present a method based on the master-slave architecture using a single camera, where scene partition strategy provides a division of the scene into several regions. As a result, the method created predefined point with specific position of PTZ. In our method we adopted the use of one camera as master function and another device as a slave to avoid camera moving and delay, resulting in more time to process frames. Instead of the scene partition, we use points provided by the targets located in any-place of scene to execute a learning algorithm, hence it is not limited by a grid or cooperative training to provide the corresponding points. In a similar approach, Chang et al. [4] introduces a master-slave to take close view of the target in a wide scene. The control system is based on estimation of the homography between cameras to provides the corresponding points in both static and PTZ camera. Due the limitation imposed by homography, the slave camera need to return the predefined position to focus in different places of the observed space. With our learned model trained in different position of the scene the slave camera is able to change the pose without to return predefined position.

Measuring the progress of online application is not a simple task, however some work has been focus in solve this problem. Chen et al. [5] describe a PTZ tracking evaluation based on a virtual camera, where spherical video sequences are captured in indoor environments with multiples person walking. With the recorded videos is possible to place the camera model inside the spherical video and evaluates PTZ control methods. Although, the system evaluates a single active camera, a master-slave setup can not be tested due the limitation of the scene and model. However the metrics TCE and TF are important to incorporates quantitative metrics in our work. In a similar approach Liu et al. [8] presents a virtual PTZ camera platform where tests can be done with different algorithms under the same environment. The main idea is providing a large panoramic background image where the PTZ camera can be simulated as a digital device. Although this method can evaluate tracking algorithms with a virtual cameras, there is a limitation of the virtual image and features such as variation in target motion and network delay. Although, previously work present a way to reproduce experiments with PTZ camera we concentrate our effort in master-slave setup, which the evaluation involves more cameras and the complexity increases.

3. Proposed Approach

The proposed method aims at providing the correspondence between master and slave cameras without performing a traditional calibration. Initially, the method records corresponding points between master and slave cameras that will be used as input in a learning-based phase to perform targets tracking in real-time.

3.1. Recording of Corresponding Points

The corresponding points are important to train a model that control the master-slave system. For that, we configure the PTZ camera to focus in moving target around the scene and record corresponding points automatically. To perform the task, the PTZ camera executes random moves searching available targets and when the target is detected [12], the method estimates PTZ parameter that centralizes the target in the PTZ camera view. To adjust the slave camera pose, we employed the works [4, 15] which, given the current focal length of the camera, estimates the pan and tilt angles, respectively $\alpha$ and $\beta$, according to

$$
\alpha = \tan\left(\frac{\Delta x}{f}\right), \quad \beta = \tan\left(\frac{\Delta y}{f}\right),
$$

where, $\Delta x$ and $\Delta y$ represent the displacement between the PTZ image center and target center in x and y axis, $f$ represents the current camera focal length. To estimate the zoom parameter, we fix a target size with the ideal dimensions for surveillance tasks [10] and capture the information of detected target (spatial location, height and width), then a linear approximation estimates the value of focal length. Images from fixed and PTZ camera are stored, as well as, target bounding box in the PTZ camera and PTZ parameters provided from VAPIX AXIS Library.

Figure 1. Test phase flow.

We then process the images from master camera recorded during the training in off-line mode to find the exactly target located in master and PTZ views. Then, we run the object detection algorithm [12] in the recorded master images to find the same target centralized in PTZ camera and in cases with more than one person, we execute the unsupervised re-identification method (more details in [9]). This strategy allows us to handle multiple persons in the surveillance space. Although the algorithm is able to perform the re-identification, the performance was not as expected, thus we also performed manual check to ensure the correctness of corresponding points. We believe with more robust methods, which handle images from different cameras such as deep learning methods, the results could be improved. To overcome this problem we only record more points with a single target.

3.2. Neural Network Model

We developed a learning method to estimate the master-slave correspondence based on a nonlinear regression where given coordinates of the target bounding box in the master camera view, we estimate the value of PTZ for the slave camera to focus that particular location. During the phase described in Section 3.1, the method provides a list of corresponding points to integrate the input and output data of a learning method. Then, we develop a fully connected neural network that receives the target bounding box location from master camera view as input and returns absolute values of PTZ parameters. This type of strategy can perform a nonlinear regression adjusting perfectly with the problem. The choice of deep-learning strategy is motivated by the large applicability and robustness for adaptation in different type of problems. We have tested linear models, but the results were not as expected.

With a model and weights created during the training of fully connected neural network it is possible to run the test. During the test phase, a fixed camera with wild view provides images to run the object detection [12] that creates target location \((x, y, w, h)\) and these information are forwarded into neural network. Then, the loaded model predicts PTZ camera parameters that centralizes the target in PTZ camera field of view and a high-resolution image is available to be processed afterwards. The Figure 1 presents the test phase flow.

4. Master-Slave Evaluation Setup

Recent studies have focused in PTZ camera evaluation. However, comparing master-slave setups there is no a standard methodology. Therefore, we present a experimental setup in which we are able to execute two different methods simultaneously and with almost the same view for each. A master camera shares the view for each method and the pair of active camera runs different methods. Thus, the evaluation is performed simultaneously considering the same scene and allows a fair comparison of the methods.

4.1. Experimental Setup

The setup is composed of two PTZ cameras positioned side-by-side in the same location and a wide view camera is the master. We install the setup in an outdoor scene, where uncontrolled factor are present, such as occlusion and illumination variance. The master camera was installed at 8.47 meters of height and the pair of PTZ camera was placed at 6.85 meter far from master in the same height. The pair of PTZ camera is placed with the displacement of 0.18 meters between each other, due physical limitation. Figure 2 illustrates the proposed experimental master-slave camera setup and the surveillance space. In addition, we measured the limitation of the field of view caused by the proposed setup and we found a reduction of twenty-three degrees, considering that the PTZ camera has the field of view of 180 degrees.

4.2. Evaluation Metrics

Evaluating master-slave setup is not simple task since it involves the use of online data that cannot to be reproduced (one cannot captures the same scene twice). Thus, we propose a master-slave setup where is possible to evaluate two methods of control simultaneously (as described in Section 4.1) and used two metrics to generate quantitative results to compare the two methods.

The first metric evaluation measures the displacement between coordinates of target center and center of PTZ camera image plane, named as Target to Center Error (TCE), as

\[
TCE = \frac{1}{n_{vf}} \sum_{j=1}^{n_{vf}} |c_v - c_t|, 
\]

where, \(n_{vf}\) represents the number of valid frames during the execution, \(c_v\) represents the image plane center and \(c_t\) is center of detected target. Finally, the second metric, Track Fragmentation (TF), evaluates whether the target is inside or outside of the active camera view.
5. Experimental Results

We implement the system in Python 3.6 and OpenCV 3.4. All tests were performed on an Intel(R) Core i5-4440 at 3.10 GHz, 16Gb RAM and a GPU NVIDIA GeForce GTX 770, using cameras AXIS P1357 with resolution $1280 \times 720$ pixels (master camera) and two PTZ cameras AXIS P5512 with resolution $704 \times 480$ pixels with pan and tilt speed of a hundred degrees per second (slave cameras).

5.1. Neural Network Architecture Parameters

We performed extensive tests, varying the number of neurons, layers, activate function and loss-function, to select a ideal architecture that best fits in the problem. The input of fully-connected neural network is a matrix of target bounding box $(x, y, w, h)$, the first layer has fifteen neurons, the second layer has ten neurons and the output layer has three neurons, which represents the PTZ parameters. The loss function adopted was the mean squared error (MSE). During the training, the number of epochs was not limited and a callback was set to stop the training whether the loss function does not decay during 50 epochs. All hidden layers are equipped with the rectification non-linearity (ReLU) activation function and the last layer uses the hyperbolic tangent activation function.

5.2. Neural network Training - Pan and Tilt

To investigate the performance of the neural network for pan and tilt parameter, we trained the network optimizing the regression objective using mini-batch gradient descent based on back-propagation and Adam parameters [6]. The batch size was set to 15, the dropout was not used and the data evaluation was 70% used in training and 30% of data applied for test. The number of epochs was not fixed and the training stopped when the loss decay does not decrease during 50 epochs.

We perform the neural network behavior predicting values of pan and tilt, which the zoom parameter was fixed in minimum value (1x). Figure 3 shows two qualitative results of our method.

5.3. Neural network Training - Pan, Tilt and Zoom

To acquire images with high-resolution, we train the neural network using the same parameters presented in the last section (Sec. 5.2) however, we introduce the zoom parameter. Thus, the neural network model will be able to predict the PTZ parameters.

First, we check the average error during the training varying the number of corresponding points recorded in the training. We perform the test with 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200 and 1300 corresponding points (Figure 4, left image). According to the results, the accuracy of the method is statistically invariant, thus we can train the neural network with a low number of points and the model is able to generalize during the execution.

To confirm the invariance between the models with different number of points during the training, we also test the performance of model during an execution in the outdoor scenario. We execute each created model and measure the error in pixels using the metric TCE in two points of the surveillance space (Figure 4, right image). In spite of the divergence between models, the results are statistically similar. Thus, we decide to choose the model trained with 1200 points.

After the choosing the model, we test two executions of the method, where camera parameters are predicted and more detail of the target are visualized at the surveillance space. Figure 5 illustrates the execution during a period of time. The left column represents the master view in time $t$, the middle column also represents the slave view in time $t$ and the right column contains the adjust of slave camera for time $t + 1$. 
5.4. Experimental Setup Evaluation

Given the defined camera setup (Section 4.1), we investigate the displacement between each camera view, as illustrated in the Figure 6. We observe a low discrepancy of each view when a simple difference was applied in both image view. The small drift observed in the second and third columns is due to the displacement between each camera and the mechanical error of zoom when the parameter of pan and tilt have been fixed.

In a detailed evaluation to demonstrate the small difference of view with the proposed setup, we test for different values of zoom (1×, 6× and 12×) and fix values of pan and tilt parameters in both camera. We detect keypoints in both views applying a feature detection and matching with SIFT method, as in Bimbo et al. [2]. After selecting matched points, we computed the Euclidean distance to find the average difference in pixels of camera views, which represents the displacement between the two active cameras. Figure 7 shows the average (in pixels) and illustrates some matched points. Even though the disparity increases when the zoom is increased, we believe that this type of model allows the creation of a mapping function between the parameters of both cameras, which should compensates the displacement error.

<table>
<thead>
<tr>
<th>Test</th>
<th>Method</th>
<th>TCE (pixels)</th>
<th>TF(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Our with PT</td>
<td>33.1 ± 21.5</td>
<td>99.4</td>
</tr>
<tr>
<td></td>
<td>Our with PTZ</td>
<td>93.9 ± 54.7</td>
<td>97.9</td>
</tr>
<tr>
<td>2</td>
<td>Our</td>
<td>79.9 ± 52.2</td>
<td>98.1</td>
</tr>
<tr>
<td></td>
<td>Our without PP</td>
<td>117.5 ± 70.6</td>
<td>82.5</td>
</tr>
<tr>
<td>3</td>
<td>Our</td>
<td>95.2 ± 49.8</td>
<td>98.8</td>
</tr>
<tr>
<td></td>
<td>Wang et al.[14]</td>
<td>143.9 ± 75.3</td>
<td>98.1</td>
</tr>
</tbody>
</table>

Table 1. Performed tests.

5.5. Results and Comparisons

To provide a comparison between camera control approaches based on master-slave, we execute using the proposed experimental setup and recorded a target walking autonomously in the surveillance space during three minutes. Then, three tests were performed. The first evaluates our method to track a target with pan, tilt and zoom parameters versus our method using only pan and tilt parameters. The second evaluates our method with variation in position prediction and the third compares our method with Wang et al. [14]. The procedure of annotation was made executing the object detection [12] and a human supervising and refining target bounding box. After that, metrics TCE and TF were computed.

We test our method using only pan and tilt parameters to verify the precision when compared with the use of pan, tilt and zoom (Table 1). According to the results, the error increases when the zoom is added, which demonstrates that a error with small angle is significant when the zoom is applied. However, more detailed images are recorded and the target can be observed. In the second test, we compare our method with the implementation using a target position prediction and without position prediction (PP). According to Table 1, the addition of a simple prediction is able to keep the target centralized in the slave camera more accurately. In the final test, we compare our proposed approach with the method proposed by Wang et al. [14]. According to Table 1, we can observe that our method achieves better results when we compare the capacity of keeping the target centralized in slave view. Our method presents the average error of 95.2 versus 143.9 pixels and also keeps the target in the slave view with a higher rate.

In general, the proposed evaluation experimental setup can evaluates different method simultaneously and provides a fair comparison. We provide a video which contains part of the three tests as Supplementary Material.

5.6. Difficulties and Limitations

The master-slave setup control provides detailed information of a target in surveillance space. Although a large
number of works have been improved the methods, some limitations restrict the performance of our method. The first limitation is related to the target detection – when it fails there is no options to execute and external factor as illumination, blurring and occlusion affects the operation negatively. To overcome this problem, a detector with greater accuracy can be used combined with a robust position prediction. Another limitation was related to the re-identification because unsupervised methods do not behave as expected for different cameras and different targets size – a supervised learning method can improve the results. Finally, incorrect detections provide poor parameters estimation which reduces the accuracy of our method – the used of historical information could reduce such problems.

6. Conclusions and Future Directions

In this paper, we proposed a new master-slave camera control that uses a neural network method to define the correspondence between fixed and PTZ cameras. Moreover, a new evaluation setup was proposed to compare different master-slave methods in the same scenario. The new setup was composed of a fixed camera with shared view and a pair of PTZ cameras, which perform two different methods at the same time. We carried out extensive experiments with the proposed method and evaluates showing more efficient when the focus is keeping the target centralized. Future works include a mapping function between two cameras and the cooperation of large number of cameras to survey a scenario.

Acknowledgments

The authors would like to thank the National Council for Scientific and Technological Development – CNPq (Grant 311053/2016-5), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17), the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project) and Petrobras (Grant 2017/00643-0).

References


