Calibration-free Camera Hand-Over for Fast and Reliable Person Tracking in Multi-Camera Setups

Birgit Möller*, Thomas Plötz† and Gernot A. Fink†
*Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Halle, Germany
Birgit.Moeller@informatik.uni-halle.de
†Intelligent Systems Group, Robotics Research Institute, Dortmund University of Technology
Dortmund, Germany
{Thomas.Ploetz, Gernot.Fink}@udo.edu

Abstract

Ensembles of multiple (active) cameras yield an important ingredient in modern tracking and surveillance applications. They overcome the limited fields-of-view of single cameras, however, require robust procedures for handing over tracking tasks from one camera to another. In this paper a calibration-free procedure is proposed that allows for fast and reliable camera hand-over in Ambient Intelligence (AmI) applications. The approach is based on online acquisition of scenario-specific target models and especially solves the problem of significant changes in object view during hand-over. Real-world results acquired in an AmI environment prove the effectiveness of our technique.

1. Introduction

For a variety of applications fast and reliable tracking of moving objects in video sequences is of essential importance [13]. Apparently, surveillance is one of the most prominent domains substantially relying on person tracking. Furthermore, especially within the domain of Ambient Intelligence (AmI) applications, e.g. for smart environments, permanent knowledge of a person's position is an important source of information.

Person tracking gets rather complicated when multiple (active) cameras need to be used. Such multi-camera setups especially become necessary when a person to be tracked leaves the (limited) area covered by some particular camera C_1. In this case, a camera hand-over towards the next suitable camera C_2 needs to be performed. If successful, this allows the global tracking procedure to be continued based on the data of camera C_2. Changes in camera views, however, represent the basic challenge for the required camera hand-over. As an example camera C_1 might cover the moving person frontally with large zoom whereas camera C_2 might capture side view images of the person to be tracked with small zoom. In order to synchronize these different views the cameras involved need to be calibrated.

Unfortunately, for various dynamic scenarios the required exact mutual calibration of the cameras appears to be unsuitable. Especially in the aforementioned AmI-domain camera positions are often changing, e.g. since they are mounted on mobile devices, or due to the need for fast re-configuration of the particular smart environments, preventing an exact mutual calibration.

Our research is, generally, concentrated on the development of a smart house – the FINCA [11]. Within the context of this paper we present a fast and reliable approach for tracking persons in video sequences which is suitable for the aforementioned dynamic scenarios. Persons are tracked using the Mean-Shift algorithm [3] based on hue color histograms. If the person to be tracked leaves the area covered by a particular camera the tracking process is seamlessly handed over to the next suitable one. In order to continue the overall tracking process on the data of the next camera (exhibiting a different view on the scene), first a rapid initialization technique is utilized providing the local starting points. For robust continuation multiple adapted color histograms extracted by the previous camera during the preceding tracking phase are then evaluated in parallel.

The main contribution of this paper is twofold. First, the proposed multi-camera tracking procedure does not rely on exact mutual calibration of the cameras involved. Second, the overall multi-camera based tracking procedure is fast and reliable due to the integrated evaluation of multiple scenario-specific color histograms. This represents an important pre-requisite for the addressed AmI-domain as we exemplarily show in an experimental study within the FINCA.

2. Related Work

The problem of handing over tracked objects from one camera to another is closely linked to the general

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

question of how to establish correspondences between different cameras that share a common field-of-view. In particular, views from different cameras need to be matched to retrieve a target object again. Correspondences are usually established, e.g., applying color- or feature-based region or shape matching techniques [4], in static setups often combined with background subtraction routines for reliable object detection. To simplify this task usually not the complete images are scanned for matching objects, but the problem complexity is reduced by incorporating additional knowledge.

In the literature the majority of approaches relies on calibration data. Given a fully calibrated 3D-space, correspondences in different views can be calculated directly without any search [5]. Applying a more weak calibration in terms of epipolar constraints or ground plane assumptions at least allows to restrict the correspondence problem to a lower dimensional subspace [1, 9]. In [2] and [8], respectively, calibration data is extracted in a more qualitative way by learning the relations between the fields-of-view of different cameras and transitions between them during a training period. However, this takes either a lot of time or requires controlled motion in the observed scene during training.

Generally, calibration is feasible for static multi-camera arrangements. The complexity of the task significantly increases if active cameras are involved. In [6] and [7] approaches for offline camera calibration of ensembles of static and dynamic cameras are proposed. However, their complexity recommends them mainly for scenarios where high accuracy is mandatory.

Moreover, even calibrated setups do not allow for an easy solution of the problem that different cameras may have significantly varying views on the tracked object at the time of hand-over. Even in suitable subspaces simple color or shape matching routines will fail in these cases given only the current information. Accordingly, in [10] an approach for active multi-camera tracking is proposed using different color models of an object that represent different views. But, these models are learned explicitly during a training phase which is not suitable for dynamic scenarios as addressed by this paper.

3. Calibration-Free Camera Hand-Over

Reconsidering the smart house project in which our developments are embedded, the general goal is to perform robust person tracking. Even people that never entered the FINCA before and, thus, are unknown to the system, should be tracked robustly so that at every point in time their position is known. To reach this goal we utilize multiple cameras each with a limited field of view. Accordingly, at certain 'transfer points' the tracking procedure of one camera needs to be handed over

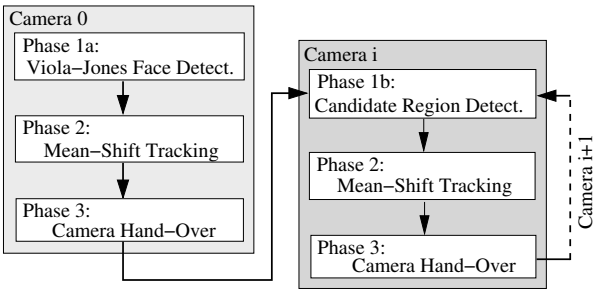

Figure 1. Overview of the approach.

to the next suitable one. Thereby our main intention, discussed here, is to overcome the need for any camera calibration procedure. We simply assume that the transfer points (or regions) are only coarsely defined just by knowing the overall geometry of the surveillance area.

We developed a three-phase tracking procedure free of geometric calibration (cf. Fig. 1). We only assume the cameras to be radiometrically calibrated, which can easily (and in principal even automatically) be done in our scenario given the known camera arrangement. The first phase (1a) of our approach corresponds to initialization and is performed once a person enters the monitored room. Following this, in the second phase (2) single camera tracking of the moving person is performed using Mean-Shift based on hue color histograms. Meanwhile a model of the tracked person consisting of a series of hue and RGB color histograms of the tracked region is created. In the third phase (3) this model is handed over to the next suitable camera (according to the coarse geometry of the area monitored) and used for setting up and continuing with the tracking on that camera (1b), *without* explicitly searching for the person again. For continuous operation this hand-over procedure is independent of a certain camera, subsequently allowing for further transitions.

3.1. Person Tracking

The entrance of our smart conference room is monitored by a dedicated camera capturing close-ups of the persons' faces entering. This specialized setup on the one hand results in a very limited field of view of this "door" camera but, on the other hand, it allows for very robust face detection. Consequently, we initialize the person tracking procedure by means of a Viola-Jones (VJ) detector [12] on door images (half PAL resolution). The first face which is detected in ≥ 2 frames at a stable position (small variance within center of rectangle) is used as starting point for tracking. To ensure robust tracking the hypothesis rectangle is, afterwards, slightly minimized. A standard Kalman filter is used for motion modeling and prediction, ensuring proper initialization of the Mean-Shift tracking in each image even in case of large object shifts between subsequent frames that

otherwise would cause the Mean-Shift tracking to fail.

Once the face region is found, Mean-Shift tracking [3] based on 32-bin hue color histograms is performed. Histogram comparison is based on the Bhattacharyya-distance (see below). The initial histogram is determined from the adapted face region provided by the VJ-detector. During tracking a scene specific model given by extracted hue and RGB histograms is created. This model serves as the basis for the robust continuation of the tracking procedure when camera changes occur.

3.2. Hand-Over based on Color Histograms

Once the tracked person reaches a transfer point between two cameras, i.e. the border of the field-of-view of the first camera, the hand-over is performed. The scene specific model acquired during the second phase of Mean-Shift tracking with the first camera, and given by the set of collected color histograms, is passed over to the second camera. Since scanning the whole image with every given histogram to search for matching positions and the person's head, respectively, is cumbersome and incompatible with real-time hand-over, a hierarchical relocation procedure is applied instead.

Initially for each RGB histogram the bin with maximum weight is used to define a range of reasonable color values for each channel that corresponds to the tracked person's main RGB color characteristics. This allows for a rough, but quite fast initial localization of candidate regions which is sufficient at this stage of the approach. Hence, thresholding the input image according to these ranges, post-processing the resulting binary image with morphological operators and subsequently labeling connected components yields a map of candidate regions. Only for those regions that exhibit a size compatible to some tolerance bounds (chosen according to the usual size of a person's facial region as visible in the view of the second camera) a more detailed similarity check is performed. In particular, the local RGB color distribution is compared to the reference histogram based on the Bhattacharyya-distance. The best region match exceeding a certain similarity threshold (≈ 0.4 to 0.6) defines the target region to be tracked.

To improve the robustness of the procedure each tracker starts in an initial verification state. This status holds for several frames during which the hypothesized correspondence is verified. Only if the tracked object shows some motion the target is approved and tracking continued. Otherwise it is canceled and the tracker reinitialized by searching for better candidate regions.

4. Experimental Results

In order to test the effectiveness of the proposed person tracking approach we conducted certain experiments within the FINCA's conference room (Fig. 2).

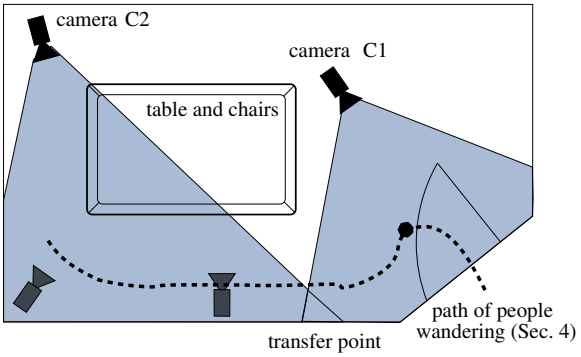

Figure 2. (Simplified) Sketch of the FINCA.

Different people were asked to enter the room and wander around unconstrainedly. Camera C_1 is dedicated to monitor the door only, while the field-of-view of camera C_2 encompasses another section of the room. Both cameras partially overlap in their fields-of-view.

The test data consists of image sequences for seven persons (of different complexions) captured from both cameras. In total 1,550 images ($C_1 = 740$, $C_2 = 810$) were recorded (378×278 pixels) at $\approx$ 6fps. After hand-over up to 50 histograms (extracted from the sequences captured by C_1) are used for searching.

We used standard hardware for both capturing (Sony Evi D70P) and computing (3.5 GHz Athlon with 1GB RAM running under Linux). The proposed tracking system was implemented in C++ as IceWing plugins (http://icewing.sourceforge.net).

During our tests the system was able to successfully track six of the seven persons during the whole particular sequences. Hence, the (sequence-based) success rate is 85.7%. Only in one sequence the hand-over failed, i.e. the tracked person was not found due to extremely poor lighting conditions, resulting in doomed tracking.

The performance of the whole procedure is satisfying w.r.t. the desired application. Face detection requires $\approx$130ms, and Mean-Shift tracking for the first part takes $\approx$50ms per frame. After hand-over candidate region search and tracking using multiple histograms takes $\approx$170ms. Performance details for tracking after hand-over are summarized in Table 1. The second column shows the number of frames required until relocation of an object, where lower counts indicate faster camera switches. On average after $\approx$ 7 frames, i.e. 1-2 sec., the tracked person is found on the particular second parts of the sequences and an average of 6 histograms match the person initially.

In Fig. 3 detailed results for tracking and hand-over are given for one exemplary sequence. In the top row a collection of images acquired with camera C_1 is shown including a person that enters the FINCA. From the rectangles in the images that indicate tracked positions

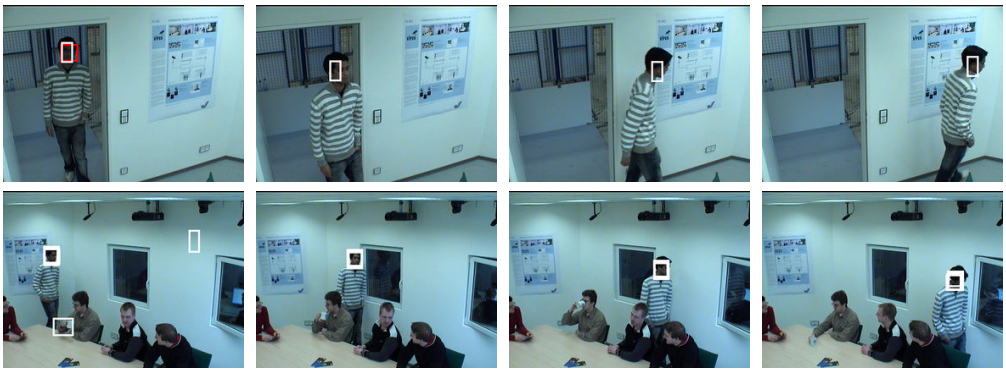

Figure 3. Tracking results (Seq. 6), 1st row C_1, 2nd row C_2 after hand-over.

Table 1. Performance after hand-over.

Sequence	# frames until person found	# histograms matching
1	9	12
2	1	3
3	12	10
4	16	6
5	1	3
6	2	1

of the head over time it can be seen that based on the initial face detection the tracking works reliably and stable. In the bottom row a subset of images acquired by the second camera C_2 is depicted, where the person enters and passes by the table where a meeting takes place. The person is robustly detected, and successfully tracked until leaving the viewing field of C_2 again.

If in some frames, recorded by the first camera, the target is only roughly located erroneous histograms are included in the scene specific model. These may cause spurious target detections by the second camera (cf. first image in second row of Fig. 3). While obvious false alarms, e.g. caused by non-moving objects, can easily be identified and corrected (Sec. 3.2), the problem is harder to tackle if there is an inherent similarity between two histograms. This may also happen if persons are dressed quite similar. For the moment these ambiguities are not explicitly resolved to prevent any loss of information. Instead competing trackers are initialized to track all candidate regions. In the future an additional verification of the different hypotheses might be carried out, e.g. by temporal integration.

5. Conclusion

We presented a calibration-free camera hand-over procedure that allows for fast and reliable person tracking in multi-camera setups. Mean-Shift tracking is applied using color histograms, which are part of a model of the tracked person. The model is created while tracking and consists of scenario specific color histograms. It is handed over from one camera to another at specific transfer points. Following this, tracking is seamlessly continued on the data of the next camera. All cameras involved are not required to be calibrated mutually. By means of experiments in a smart conference room we demonstrated the effectiveness of the new approach.

References

[1] J. Black, T. Ellis, and P. Rosin. Multi view image surveillance and tracking. In *Proc. of Workshop on Motion and Video Computing*, pages 169–174, 2002.

[2] S. Calderara et al. Consistent labeling for multi-camera object tracking. In *Image Analysis and Processing*, volume 3617 of *LNCS*, pages 1206–1214, 2005.

[3] D. Comaniciu et al. Real-time tracking of non-rigid objects using mean shift. In *Proc. IEEE CVPR*, volume 2, pages 142 – 149, 2000.

[4] O. D. Faugeras. *Three-dimensional computer vision: a geometric viewpoint*. MIT Press, Cambridge, 1993.

[5] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[6] R. Horaud, D. Knossow, and M. Michaelis. Camera cooperation for achieving visual attention. *Mach. Vision Appl.*, 16(6):331–342, 2006.

[7] A. Jain et al. Using stationary-dynamic camera assemblies for wide-area video surveillance and selective attention. In *Proc. IEEE CVPR*, pages 537–544, 2006.

[8] Y. Jo and J. Han. A new approach to camera hand-off without camera calibration for the general scene with non-planar ground. pages 195–202, 2006.

[9] A. Mittal and L. Davis. Unified multi-camera detection and tracking using region-matching. In *Proc. IEEE Workshop on Multi-Object Tracking*, page 3, 2001.

[10] K. Nummiaro et al. Color-based object tracking in multi-camera environments. In *Pattern Recognition, Proc. DAGM*, LNCS, pages 591–599, 2003.

[11] T. Plötz. The FINCA: A Flexible, Intelligent eNvironment with Computational Augmentation. http://www.finca.irf.de, 2007.

[12] P. Viola and M. J. Jones. Robust real-time object detection. Technical Report CRL 2001/01, Cambridge Research Laboratory, February 2001.

[13] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13, 2006.