Counting People from Multiple Cameras

Vera Kettnaker Ramin Zabih
Cornell University
Ithaca, NY 14853
kettnake,rdz@cs.cornell.edu

Abstract

We are interested in the content analysis of video from a collection of spatially distant cameras viewing a single environment. We address the task of counting the number of different people walking through such an environment, which requires that the system can identify which observations from different cameras show the same person. Our system achieves this by combining visual appearance matching with mutual content constraints between the cameras. We present results from a system with four very different camera views that counts people walking through and around a research lab.

1 Introduction

In the past few years cameras and digitizers have become widely available, which has created a critical need for tools to analyze the content of digital imagery. Several well-known journals have devoted special issues to this topic (including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Computer*, and *Communications of the ACM*) but there has been little work on content-based analysis of video from multiple cameras. The most prominent advantage of using multiple cameras is that it becomes possible to cover spatially extended or cluttered environments, such as the inside of buildings.

One of the tasks that arise in multi-camera systems is that of counting the number of different people that move through the environment. This problem could not be solved by counting the people present in each video stream independently and adding them up, because some people would be counted several times as they move around. This task can only be solved by analyzing the content of multiple camera streams together. It is especially important to automate such multi-camera content analysis tasks because it is very difficult for humans to attend to multiple video streams at the same time.

In this paper we address a multiple camera setup where the cameras are so far apart that their view fields do not overlap, because we assume that cameras with overlapping view fields can be treated as a single virtual camera. In such an environment it becomes difficult for a program to determine how many different people were seen, because views of the same person from different cameras can be quite dissimilar from a machine point of view. This precludes the use of traditional Compute Vision techniques for matching images (such as motion or stereo), which assume that the views are very similar.

Our principal innovation lies in the use of contextual constraints to help solve the problem of recognition between different camera views . In a very general class of multi-camera setups, the content of video in each camera strongly constrains the content of the other camera streams. These mutual constraints can facilitate the content analysis of the video streams considerably. Consider the top of figure 1, where Rose starts to walk through an office environment. When she disappears from camera 1, we can infer from the floor plan that she will reappear next in either camera 2 or 4, but not in camera 3 because she will be seen in camera 2 before she can reach the location of camera 3. The bottom of the figure illustrates the problem we try to solve: people walking past cameras lead to observations, ordered in time. The system's task is to link those observations that show the same person. To achieve this task, we consider not only the floor topology, but also global consistency constraints and the temporal dependency of successive camera appearances of people walking along corridors.

In the following section we will describe how a variety of mutual content constraints can be exploited to facilitate the content analysis of multi-camera video, and how a solution can be found in an efficient manner. We will give a short survey of related work in section 4, before we present some experimental data from a four camera implementation that counts people walking through and around a lab. We conclude by discussing some future work.

0-7695-0253-9/99 $10.00 © 1999 IEEE

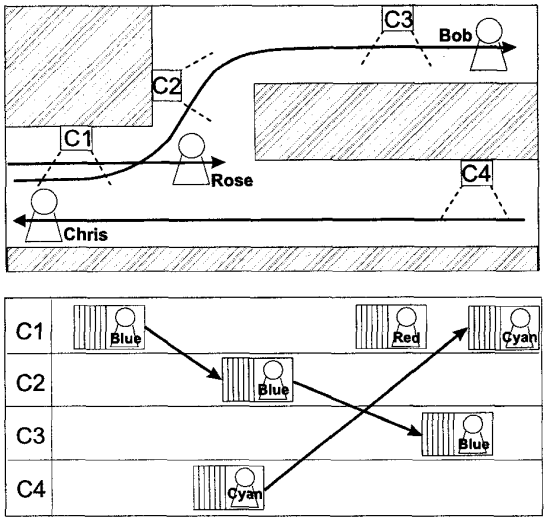

Figure 1: Top: A schematic tracking situation with 4 cameras. The arrows depict the walking trail of each person. Bottom: the resulting tracking intervals, assuming that Rose wears red, Bob blue, and Chris cyan. Time increases from left to right, and the arrows are the correct links that our system is supposed to reconstruct.

2 Exploiting content constraints

By modeling the relationship between the different cameras video streams, we can significantly simplify the problem of counting people.

2.1 Anticipated changes in appearance

Our approach begins by tracking people as long as they are visible in a single camera. We detect and segment moving objects by using a background subtraction scheme followed by a variant of Boykov, Veksler and Zabih's [2] semi-global algorithm that segments an image into maximal regions of pixels that look and move similarly to their neighboring pixels. This usually has the effect that background pixels with only a light shadow or reflection of the pedestrian are grouped with the (stationary) background, and are therefore removed from the area of moving pixels.[1] This leads to a collection of frame cutouts for each short interval of time during which a person is visible in a camera. In the following, we will refer to such a collection of frame cutouts as an observation interval. The core task in the people-counting problem is that of determining whether two observation intervals resulted from the same person. In principle, people's visual appearance changes little in our task, since people tend not to change their clothing in corridors. Our system learns a statistical model of visual appearance for each observation interval. This model is then used to determine if future observation intervals are likely to stem from the same person as the observation interval for which the model was constructed. Unfortunately, a person can look very different for a machine if viewed from different cameras, due to changing light conditions, differing camera angles, and partial occlusions. Instead of directly modeling how the appearance of objects will change between each pair of cameras, we opted for the computationally more efficient solution of normalizing the lighting conditions of all cameras with respect to a single reference camera and using a representation of visual appearance that is inherently robust to many of the anticipated changes. We currently histogram the pixels of each frame cutout by table lookup into a 32-bin color space after normalizing the lighting. Our color space was designed to capture the subtle difference between popular clothing colors such as beige and gray, while not making overly fine distinctions of colors such as red that span a wide range of the HSV color space but are perceived as rather uniform by humans. We then quantize the counts in each histogram bin into chunks of one fifth of the blob area and model the number of such chunks per histogram bin as poisson-distributed variables. This ensures invariance to changes in size of the blob area and makes the representation quite robust with respect to changes in viewing angle.

[1] Note that reflections and shadows are difficult as they move at the same speed as the pedestrians.

2.2 Reoccurrence locations and times

If it would be feasible to completely normalize visual appearance (as humans seem able to do), one could solve the task of counting people by grouping together all those views that show the same person, without considering at which location and time each frame was recorded. However, current machine recognition technology is not yet robust to changes in viewing angle for objects that were learned on the fly under rather uncontrolled conditions. Therefore, the recognition performance can be improved if other constraints are used to limit the number of potential match candidates. Our system exploits knowledge of the corridor topology and usual walking speed of people to form expectations about the time windows and the locations in which people will reappear next. Most corridors only lead to limited numbers of other corridors, which means that oftentimes, the resulting transition probabilities alone already constrain strongly in which camera streams a person could appear next. If

one makes the simplifying assumption that people's decision on where to go next from a location does not depend on where they came from, then the location and time of reoccurrence of each person can be modeled as a semi-Markov model [5]. These 'soft' constraints serve as Bayesian priors for the recognition.

It is important to note that a solution for all observations cannot be simply found by determining the best successing appearance for each observation, since this would most probably lead to inconsistent solutions, for example solutions that state that an observed person was seen at multiple locations at once. We have derived an a posteriori expression for the probability that a certain grouping of all observations into views of one person each is the correct explanation of all observations. The approach and derivation is an extension and modification of a probabilistic formulation of a radar tracking problem [7] due to Poore. The expression considers other factors as well, such as the frequency with which new people enter the environment at different locations.[2] An optimal solution can then be found by maximizing this probability expression. The following section will explain how to efficiently optimize this probability in a way that ensures that only consistent solutions (i.e., mutually exclusive chains) are considered.

2.3 Mutual exclusion of content

Each person walking through the environment will cause a chain of observations in the different cameras, and these chains will not branch if we assume non-overlapping cameras, because each person can only be in one place at a time. Fortunately, it is possible to determine that set of non-overlapping, non-branching chains of observations which has the highest a posteriori probability without having to search through all possible sets of observation chains. In order to find such an optimal set of chains in a computationally efficient way, we do not maximize the above mentioned a posteriori expression directly. Instead, we maximize the relative improvement of a hypothesis with respect to a reference hypothesis which states that all observations stem from different people. By making a few simplifying assumptions and applying some algebraic transformations, it is then possible to arrive at a set of terms each of which expresses the goodness of a hypothesized chain link. For these link terms, it holds that maximizing their product leads to solutions that are optimal under the assumptions we made. The main assumptions that our approach makes are that people's movement can be modeled as a semi-Markov process and that visual similarity is transitive, in the sense that we only consider how similar an observation is to the observation that is hypothesized to be its immediate successor, and not to any previous observations. It further assumes that the visual appearance of a person is independent of other people's appearance or behavior and that the observations intervals returned by the low-level tracking show all people that pass a camera, and only one person per observation interval. Maximizing the link term product will lead to a solution that observes the constraints on the visual appearance and spatio-temporal reoccurrence of successive observations. The additional mutual exclusivity constraints are enforced by treating this constrained optimization problem as a weighted assignment problem, which is a special case of a linear program. Weighted assignment problems can be solved optimally and efficiently, for example by the Munkres algorithm [3] which we currently use to find a solution.

[2] Due to spatial limitations, we refer the reader to a forthcoming technical report for details of the derivation and mathematical formulation.

3 Handling entrances and exits

In order to handle people entering the environment, we add a virtual observation 'NEW' that will be assigned as preceding occurrence if the best explanation of an observation is that it belongs to a person who just entered. The solution of the linear program represents exits by simply not linking any later observation to the last occurrence of a person. The probability expression for the a priori probability of a set of chains from which we derive the linear program contains terms that regulate the number of people entering the system and the typical length of a chain of observations.

Incidentally, the explicit modeling of new persons entering the environment also allows us to limit the number of previous observations that have to be considered as potential match candidates: if it is already clear from the space and time relation of two observations that they are less likely to be immediate successors of each other than the hypothesis that the current observation stems from a new object, then the two observations' visual appearances do not even have to be compared. This intuition can also be proven in terms of probabilities. For each camera pair, it leads to a (relative) time window in the past that specifies in which time range previous observations at the first camera have to lie if they can at all be the immediately preceding occurrence of an observation in the second camera. Finally, if we store the observations of each camera ordered by time, then these time win-

dows allow us to retrieve the set of potential match candidates very efficiently.

4 Related work

Boyd *et al.* [1] presented an architecture designed for multiple sensors observing a dynamically changing environment. However, their cameras overlap and the view fields are transformed into one view field via standard techniques. Although their architecture is quite general, it seems difficult to apply it to tasks such as ours where observed objects are invisible for extended periods of time.

Grimson's research group [4] has built a multi-camera system that also assumes overlapping camera fields: he envisions observing activities by a set of cameras that are strewn out in an environment and that determine automatically how to map their local view fields into one coherent view field. They then learn classes of observed behavior.

Finally, Huang and Russell [6] have designed a system that performs a task similar to ours: they monitor a highway at two consecutive locations and try to find matching cars in order to count the number of cars and to measure link times. They concentrate on appearance constraints. They also transform their problem into a weighted assignment problem, although with different link weights and problem structure, since they start from different premises than our derivation. Their solution is confined to setups where cameras are placed alongside a single path so that the movement of the objects is deterministic, with the exception of objects entering and exiting the environment. Our solution is much more general by allowing arbitrary corridor systems in which pedestrians can choose paths. Therefore, our system is able to reconstruct the paths of all objects through an environment, which is interesting for some tasks. For example, traffic planners might want to optimize traffic light controls such that traffic flow is least interrupted for the most popular routes through a city.

5 Experimental evaluation

In order to evaluate our prototype, we set up a small surveillance system of 4 cameras in and around our Systems Research Lab. The floor plan is depicted in figure 2, together with background snapshots from the 4 cameras. We conducted an experiment of about 8 minutes, where two subjects walked separately and together as many paths through the system as they could think of, always changing clothes in between different paths so as to impersonate different people. Since the experiment was conducted on a summer morning, only three additional people walked through

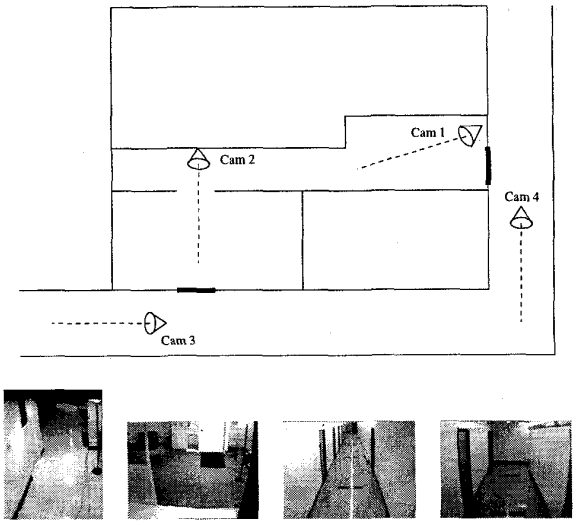

Figure 2: Floor plan of the camera setup and background snapshots from the 4 cameras.

our setup. The experiment resulted in a total of 28 observation intervals from 14 true tracks. We count the tracks that two people walked together as one track because the basic tracker consistently merged the two people together and therefore also into one observation interval. Figure 3 shows observation intervals and the correct observation links from a subsequence of the experiment. Our initial results are quite promising: only two observation intervals were linked incorrectly, and one link was broken as a result of one of the wrong linkages. In both cases of the wrong additional links, the transition times of the suggested links were likely, and the clothing of the correct and wrong matches had similar color and differed only in the pants' length. However, the data also contains two cases in which the same person appears twice after an unnaturally long disappearance time (thus violating the modeling assumption that people would just walk along the corridors). In these cases, the system labels the second appearance as 'NEW'. Taken together however, the system counts 13 different people, which compares well to the 14 true tracks.[3]

These first results were obtained in difficult lighting situations and with a very weak representation of visual appearance, as well as significant segmentation errors.[4] But they nonetheless suggest that our

[3]We obtain this number by forgiving the system for marking reappearances after unnaturally long disappearance times as new people. Without this assumption, there are 12 true tracks.

[4]The segmentation errors were due to strong reflections and shadows on the hallway floor.

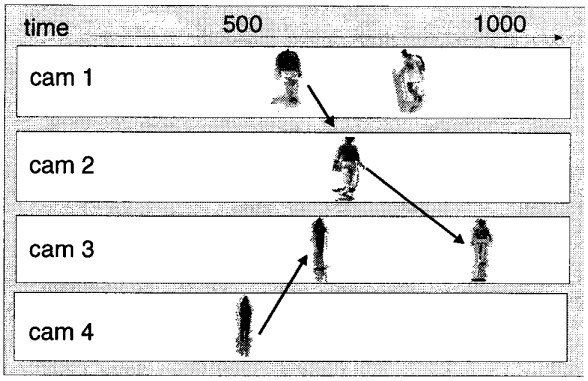

Figure 3: An example subsequence of the experimental 8 min sequence. Observation intervals are represented by the observation in the middle of the interval.

approach performs well and is quite robust. Our focus sets led to reasonable time windows and ensured that each observation only had to be compared with a very limited number of other observations. The average size of the focus sets in this experiment was 1.6, while without the focus sets we would have needed to compare an observation with an average of 13.5 other observations.

6 Future work

At present, the parameters for all the probability density functions, usually frequencies, were estimated from experimental data and then given to the system as input. However, it is very straightforward to update these parameters at runtime by reestimating over a limited time window, or simply by using sliding averages that discount past data points exponentially.

It is also easy to think of additional constraints that could be exploited, such as the direction of a person's movement when visible in one camera. This information could be used to discredit hypotheses that include U-turns: one could split each camera location into several virtual locations that are distinguished by the different dominant movement directions of objects in this location. Each of these virtual locations would then have different transition probabilities which would make U-turns unlikely.

We are also interested in applying our framework to other tasks and environments, such as car traffic. Our system should perform quite well for car traffic since there is more structure in the movement of cars than in the somewhat random behavior of pedestrians.

7 Conclusion

In this paper we described how mutual content constraints of multiple camera streams can be exploited to solve tasks that would otherwise be difficult. We proposed technical solutions for several different types of mutual content constraints and demonstrated the viability of our approach with a system that counts the number of different people appearing in any of multiple cameras viewing an indoor environment.

Acknowledgements

We would like to thank Carlos Saavedra for helping out as one of the subjects in the experiment. This research has been supported by a grant from Microsoft.

References

[1] J. Boyd, E. Hunter, P. Kelly, L. Tai, C. Phillips, and R. Jain. MPI-video infrastructure for dynamic environments. In *IEEE Conference on Multimedia Computing and Systems*, pages 249–254, 1998.

[2] Yuri Boykov, Olga Veksler, and Ramin Zabih. A variable window approach to early vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1283–1294, December 1998. An earlier version of this work appeared in *CVPR '97*.

[3] F. Burgeois and J.-C. Lasalle. An extension of the Munkres algorithm for the assignment problem to rectangular matrices. *Communications of the ACM*, 14:802–806, 1971.

[4] W.E.L. Grimson, C. Stauffer, R. Romano, and L. Lee. Using adaptive tracking to classify and monitor activities in a site. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 22–29, 1998.

[5] R.A. Howard. *Dynamic Probabilistic Systems*. Wiley, 1971.

[6] Tim Huang and Stuart Russell. Object identification: a Bayesian analysis with application to traffic surveillance. *Artificial Intelligence*, 103:1–17, 1998.

[7] A.B. Poore. Multidimensional assignment formulation of data association problems arising from multitarget and multisensor tracking. *Computational Optimization and Applications*, 3:27–57, 1994.