Automated Feature Registration for Robust Tracking Methods

Shawn Arseneau and Jeremy R. Cooperstock
Centre for Intelligent Machines
Montréal, Québec, Canada, H3A 2A7
{arseneau | jer}@cim.mcgill.ca

Abstract

Tracking people within a scene has been a long-standing challenge in the field of computer vision. A common approach involves matching the background against the incoming video stream, with the assumption that any unmatched pixels belong to the people being tracked. Such methods, however, seem intrinsically flawed, as they do not incorporate any specific characteristics of the target in question, such as motion or shape and their performance tends be both limited and contingent upon a semi-static background. To overcome these deficiencies, we propose a saliency-based approach, which requires minimal a priori information concerning the target. Motion characteristics dictate a saliency map and highly salient regions contribute to the automated acquisition of target-specific features. In addition to improved robustness, the algorithm offers the advantages of independence from a background model and requires no explicit interaction with the user, nor imposes any restrictions on the target.

1 Introduction

Robust tracking in unconstrained environments remains a daunting challenge for computer vision. Isolating a target from the scene has typically been achieved by imposing restrictions on the environment, the target, or both. A common simplifying assumption is that the scene without the target, i.e. the *background model*, is strictly static, or contains only small dynamic locations. The model may adapt to background changes over time using temporal decay, mixture of Gaussians or non-parametric approaches. [4][9]

While the assumption of a static background may be reasonable for stationary cameras, this prohibits the use of pan-tilt units, which are, for obvious reasons, highly desirable in tracking applications or for mobile robotics. Nicolescu and Medioni, proposed a global background removal approach for such units, in which a panoramic view of the scene is stored, and segmentation is performed making use of this information, along with the camera's telemetry [8]. However, this requires very accurate telemetry or the adoption of a neighborhood approach to background differencing. Otherwise, any error will accumulate over time and eventually render the results invalid.

Background removal schemes are based on the assumption that by eliminating those parts of the images similar to the background model, whatever remains must belong to the target being tracked. Our criticism of this approach is that the emphasis is placed on matching image components to the background, rather than to the target itself. Thus, these schemes ignore potentially valuable target specific features, which could increase robustness and relax the constraints imposed on the environment. Following a more biologically motivated approach, we suggest that feature-based tracking may be more useful when applied to the features of the target itself, rather than the background.

The challenge then is to determine appropriate target features that can be tracked over a video sequence. One such feature that has been investigated is color [5]. Using one of several possible color space transformations, various researchers have attempted to identify skin colored pixels in the scene in an attempt to locate human targets. Although encouraging results have been obtained, there remain numerous scenarios in which this approach fails, for example, if only small portions of the individual's skin remain unoccluded or in situations where background objects exhibit colors within the range of normal human skin tones. Another possible target feature to adopt is object shape [2]. Given its robustness to lighting changes due to the gradient nature of the relevant calculation, shape offers great promise for tracking applications. There are, however, a number of disadvantages with shape-based tracking, primarily, the computational cost associated with such techniques. Furthermore, shape-based algorithms often assume that the initial target location is known, a prohibitive constraint for many applications [2]. Worse still, targets may exhibit non-rigid motion characteristics, as produced, for example, by the two-dimensional silhouette of a person walking. Analysis of non-rigid shapes imposes even greater computational demands on a feature-based tracking algorithm. One of the more popular features often associated with tracking techniques is motion, using such methods as optical flow or temporal

templates [1][3]. A successful person tracking scheme can be formulated by classifying observed motion patterns as related to human motion characteristics [1]. While most motion-based schemes do not necessarily segment a full outline of the person being tracked, they do indicate pertinent locations within the scene and provide a satisfactory confidence measure as to where the target may be.

Combining the results of several different feature-trackers in a multimodal fashion has been shown to increase the overall robustness of the algorithm [6]. Promising results have been obtained in this manner, employing cues such as motion, spatial characteristics and pixel intensity values [2][7]. However, significant restrictions remain, either on the subject being tracked or the environment, in order for the algorithms to provide optimal performance. The major drawback is that the target's motion, shape, texture and color model must be known a priori. In many applications however, this is not feasible as in the general person-tracking paradigm. Most people may share a common shape or motion model; however, it is unlikely that they share similar textural features. What is needed is a technique to extract such additional features on-line, which is the goal of this research.

2 Objectives

For a test case, we consider the task of person tracking without prior registration of the target in a complex, real world setting, specifically, that of a classroom environment. Our goal is to locate and track an instructor throughout a lecture, using a pan-tilt camera to keep the target approximately centered in the frame, similar to the needs for television production. This environment poses several challenges, including a wide range of lighting variations due to natural light entering through windows, and the use of LCD projectors, as well as the many instances of partial occlusion.

The tracking algorithm begins with no knowledge of the environment prior to its activation and places no restrictions on the instructor's clothing nor requires any particular identifier such as an active badge. Most importantly, the algorithm is intended to run completely autonomously, so no user initialization is involved.

3 Tracking Algorithm

As an improvement over previous work, we propose a paradigm for target tracking with minimal *a priori* knowledge regarding the associated target features. Initially, an estimate of pertinent motion in the scene is calculated. Next, regions are eliminated that do not conform to a human-motion model. [1] It is these remaining regions from which the new target features are extracted. The use of these features allows us to relax many constraints normally imposed on the background. By associating confidence values with each of the feature detectors, the algorithm can make informed decisions regarding the target location, without needing to rely solely on any one of these. It is important, however, to note that the ideal feature set is normally target specific, and thus, the selection of which features to extract remains an implementation issue.

The algorithm is divided into two main components: (1) finding an initial estimate of the target location based solely on motion cues and (2) identifying areas that correspond to a human-motion model. These areas may then be used to interpolate additional, target-specific features, such as color, texture, shape and motion patterns. These two components are discussed in the following sections.

3.1 Initial Estimation of User Location

As with any tracking system, some initial discerning feature must be known in order to distinguish the object of interest from the rest of the scene. In our case, we use motion cues from the image sequence. As one of the requirements of the tracking system is to operate in real-time, computational efficiency became of prime concern. Thus, simpler methods were favored over more computationally intensive approaches such as optical flow. Another concern is that the pixels identified as having motion should also correspond with the target itself as many motion estimation methods depict both the occlusion and disocclusion areas that do not simultaneously depict the target's present location.

One of the simpler techniques to identify regions of motion within an image sequence is through temporal differencing [3], which reflects the change between two successive frames. Although it represents change in the scene with minimal calculation, it suffers from severe ghosting on the trailing edge of the target's motion thus corrupting the data. (See figure 1) Another similar method, known as the motion history image, (MHI), represents the spatial recency of the motion in the scene. The MHI is calculated by combining temporal differencing information over time through the use of a temporal decay. This produces an image in which pixels that have not exhibited recent motion gradually decrease in intensity. This effect can be seen in Figure 1a, with pixel intensities ranging from white (corresponding to τ, the maximum value), and decreasing through darker values down to black (corresponding to 0) as the values are decayed. Thresholding the MHI by τ can be used to reproduce the most recent temporal difference image.

Figure 1b, in which the MHI is used to mask the original image, illustrates a user present within the

thresholded region; however, a ghost image trails the user, thus corrupting the data. To minimize the effect of ghosting, an additional condition, MHI(*t-1*) = 0, was added to the equation as follows:

$$MHI(t) = \begin{cases} \tau & \text{if } D_{t(i,j)} = 1, \\ & \text{and } MHI(t\text{-}1) = 0 \\ \text{Max}(0, MHI(t\text{-}1) - 1) & \text{otherwise} \end{cases} \quad (3)$$

This constraint ensures that pixels are assigned a value of τ only if they represent motion that is new to the scene, as shown in Figure 1c. The result of thresholding the modified MHI and masking it with the image at time *t* significantly reduces the ghosting effect, as can be seen in Figure 1d.

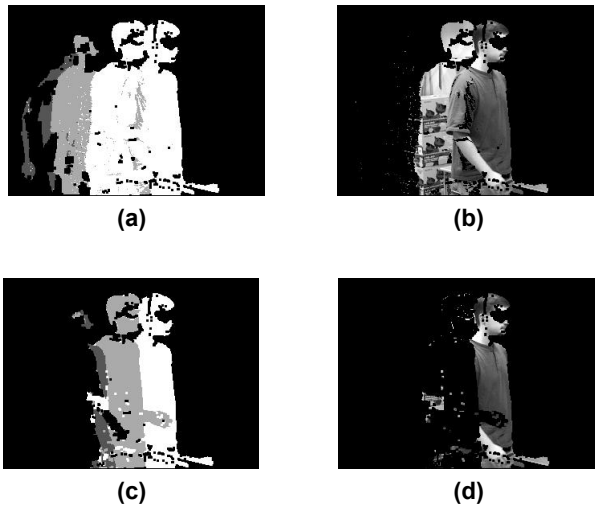


Figure 1: Partial user segmentation: (a) the original MHI, (b) image (a) thresholded with original image, (c) modified MHI with additional constraint, (d) image (c) thresholded with original image.

It is important to note that the goal of this research is not to directly segment the target from the scene, rather to identify pixels in the scene that are *most likely* to correspond to the target in question. It should also be mentioned that the resulting motion relies heavily on the frame rate of the algorithm, as well as the velocity at which the user moves. However, these parameters will be incorporated into the additional feature registration step to maximize the robustness of the overall technique.

3.2 Additional Feature Registration

Now that pixels indicating change have been revealed, they must be further processed to isolate those that more likely represent the desired target to be tracked. As mentioned before, one must have at least some a priori information concerning the target to properly isolate the appropriate region(s). To discern the motion data related to human movement, a human motion model is applied to the connected regions from the modified MHI. [1] This model is based on the regions' size and shape characteristics and has been successfully implemented as a person-tracking algorithm [1]. This stage helps distinguish candidate regions arising from dynamic components of the background (e.g. a projection screen being lowered or a door opened) from the human target. With the proper regions being isolated, any number of features may now be registered pertaining to the individual targets such as color or texture. Thus, this automated process is able to extract additional information about the target given minimal a priori knowledge about the target itself.

4 Results

The ability of the modified MHI to properly isolate pixels directly corresponding to the target was tested against three image sequences. The first sequence was of a person sitting in front of the camera and moving their upper body about in a discontinuous fashion. The second was of a person walking through the scene and the third was similar to the second with the addition of varying lighting conditions. To properly assess whether the pixels truly corresponded to the person in the scene, segmentation was performed manually on one of the frames so as to provide an accurate benchmark. Figure 2 shows the results of the temporal differencing, MHI and the modified MHI images resulting from the second image sequence. Although the original MHI encapsulates a higher number of user identified pixels than our modified method, the latter algorithm demonstrates substantially improved accuracy with respect to the percentage of correctly classified pixels. The use of a modified MHI increases the robustness of finding target pixels such that additional characteristics related to the target can be inferred and aid in tracking. This means that tracking algorithms need no longer be required to make use of a background model of the scene. Provided that some characteristic of the target is known, permitting initial location in the scene, it is possible to extract other features automatically, increasing the confidence of the tracking algorithm.

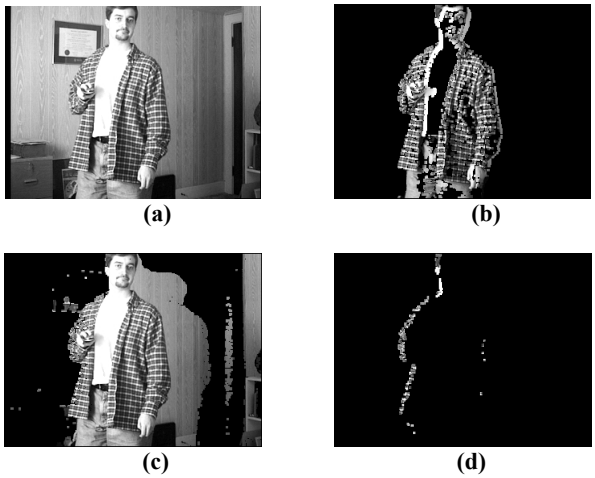
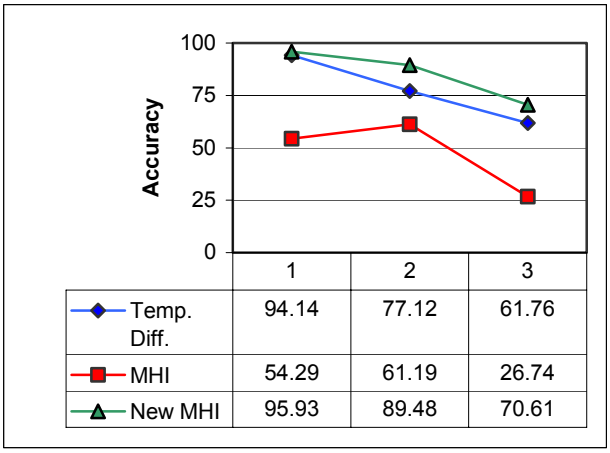


Figure 2: (a) Sample frame from seq. two (b) temporal differencing, (c) original MHI and (d) modified MHI.

Table 1. Results from the 3 Image Sequences



5 Conclusions and Future Work

The development of a tracking system that imposes fewer restrictions on both the environment and the individual being tracked significantly expands its possible application domain. Our results demonstrate the possibility of target tracking within noisy, dynamic backgrounds, without requiring manual initialization. Provided the existence of at least one distinguishing characteristic of the target is known *a priori*, additional features may be extracted during operation to increase overall robustness.

Given the unpredictability and dynamic nature of real-world environments, updating the feature detectors over time is well in order. For example, the instructor's shirt color may shift due to changing lighting conditions. Recognition of such changes could be identified using other target features, such as motion, and used to update the color detector as needed. Another useful addition would be the automatic selection of appropriate features, perhaps using Kalman filtering or a winner-take-all approach, in the event of conflicting information returned from multiple feature detectors.

Acknowledgements

Support for this research has come from Fonds pour la Formation de Chercheurs et l'Aide a la Recherche (FCAR), the Natural Sciences and Engineering Research Council of Canada, and the Canadian Foundation for Innovation. This support is gratefully acknowledged.

6 References

[1] S. Arseneau, and J. Cooperstock, "Presenter Tracking in a Classroom Environment," *IEEE Industrial Electronics Conference*, pp 145-148, 1999.

[2] S. Birchfield, "Elliptical Head Tracking Using Intensity Gradients and Color Histograms," *IEEE Computer Vision and Pattern Recognition*, pp 232-237, 1998.

[3] A. Bobick, and J. Davis, "The Recognition of Human Movement Using Temporal Templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 23, issue 3, pp 257-267, 2001.

[4] I. Haritaoglu, D. Harwood, and L. Davis, "W^4: Real-time Surveillance of People and Their Activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 22, issue 8, pp 809-830, 2000.

[5] S. Kawato, and J. Ohya, "Automatic Skin-color Distribution Extraction for Face Detection and Tracking," *WCCC International Conference on Signal Processing*, Vol 2, pp 1415-1418, 2000.

[6] S. Kim, and H. Kim, "Face Detection Using Multi-modal Information," *IEEE International Conference on Automatic Face and Gesture Recognition,* pp 14-19, 2000.

[7] I. Kompatsiaris, and M. Strintzis, "Spatiotemporal Segmentation and Tracking of Objects for Visualization of Videoconference Image Sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 10, no. 8, pp 1388 – 1402, 2000.

[8] M. Nicolescu, and G. Medioni, "Globeall: Panoramic Video for an Intelligent Room," *IEEE International Conference on Pattern Recognition*, Vol 1, pp 823-826, 2000.

[9] C. Stauffer, and W. Grimson, "Learning Patterns of Activity Using Real-time Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 22, issue 8, pp 747-757, 2000.