Presenter Tracking in a Classroom Environment

Shawn Arseneau and Jeremy R. Cooperstock
Center for Intelligent Machines, McGill University
3480 University Street
Montreal QC H3A 2A7
{arseneau|jer}@cim.mcgill.ca

Abstract

As intelligent environments become more focused on supporting human activity, we see a growing need for the computer to ascertain users' location and activity, be it for face detection, gesture recognition, or even simple tracking. To support such applications, a robust tracking algorithm is required. A camera tracking algorithm for a pan and tilt camera is proposed, that is robust to moderately dynamic backgrounds and varying lighting conditions, while requiring only modest computational resources. The algorithm is based on pre-filtering as well as temporal differencing, and achieves impressive results in the real-world setting of a university classroom.

Introduction

A fundamental component of many intelligent environments is a robust vision system that extracts essential information about a user in the scene. Whether it is face detection [3], gesture recognition [7], action recognition [4], or gaze detection [8], the video sequence must encompass the user being tracked, while minimizing the amount of noise and preprocessing.

Many current tracking algorithms are faced with the same problems as those seen in traditional image processing. These include determining *what* to track, extracting only the information needed, and addressing various forms of noise. Some algorithms attempt to achieve robustness by requiring special clothing or equipment such as data gloves [8], portable computers, or active badges [5], while others involve an initial calibration phase to identify the object of interest. However, if the goal of an intelligent environment is to support human activity, it is critical that these systems not impede users' tasks physically, nor introduce additional GUIs or setup steps, in order for them to operate. This is all the more imperative as intelligent environments are taken out of the laboratory setting and into the real world, where users are (rightfully) far less tolerant of physical encumbrances and unnecessary computer interaction.

This background motivates us to consider robust video tracking systems that can be calibrated without user involvement. One solution is to use a multitude of cameras to capture every possible location the user may occupy. However, this requires additional hardware and an unwieldy amount of image preprocessing. Since system resources are normally at a premium, an alternative is to substitute a pan and tilt camera that tracks the user, thus reducing the hardware, as well as the preprocessing requirements. This allows the allocation of computational resources to the more intensive, high-level vision algorithms that extract salient information from the video sequence.

As part of our research efforts are directed at tracking users in a classroom setting, we are particularly interested in algorithms that do so robustly, without an inordinate amount of computer resources. This particular setting poses a number of challenging obstacles including fluctuating light intensity, occlusion, and a highly dynamic background. Meeting this challenge requires an algorithm with minimal need for specific environment parameters and which functions solely on the basis of standard video input.

In the remainder of this paper, we present such an algorithm. First, some of the previous research in this area is surveyed. Next, details of the algorithm are provided, and finally, the results of this approach are discussed.

Previous Work

Background removal is perhaps the simplest and most commonly used technique to extract a person from an image [7]. This method requires that the background is either a predetermined, uniformly textured or colored area (e.g. chroma keying), or that a single, background *reference image* is available. Unfortunately, these requirements are rarely

achievable in ordinary settings. To complicate matters, background removal algorithms tend to rely on high frequency edge transitions and constant lighting levels, and further assume a static, unchanging background, in order to function effectively.

The method of the background primal sketch [9] was proposed by Yang and Levine as a technique to reduce the effects of varying light intensity. Their algorithm uses a different threshold value for each pixel in the image, based on the reflective characteristics over a sequence of images. The result is a higher threshold for pixels of fluctuating light intensity. However, changes in the background beyond the scope of the constructed primal sketch results in performance degradation due to ghosting effects.

Another popular technique is that of color matching. The image is scanned for a specific color such as skin tone [3] or clothing color. However, despite the progress of intelligent color identification techniques [6], this method is not particularly robust in that there is no universal skin tone. Furthermore, algorithms that are based solely on color can mistakenly track the wrong parts of the body. For example, a head tracker could lock onto a person's hand if it briefly occludes the face.

A related approach involves searching for a given template, identifying the object of interest [2]. This too, has its limitations, as the swatch being tracked may become occluded momentarily. For example, a face template may fail to match any part of the image when the user turns away from the camera.

In a classroom setting, the background is generally less forgiving than that found under laboratory conditions. There may be a variety of shapes and colors that are similar to the user, thus precluding the use of color tracking methods. Furthermore, one cannot rely on most presenters to "face the camera" continuously during the course of a lecture.

A more recent technique developed to overcome these problems is that of temporal differencing [4]. The difference image of scene (t+1) from scene (t) is taken, thus producing an image reflecting the change between the two scenes. The difference image is constructed as follows: Let $I_{t(i, j)}$ represent image t at pixel location (i, j), and T represent the threshold for all pixels. This produces a difference image $D_{t+1(i, j)}$ as directed by equation (1.1).

$$D_{t+1(Ii\,j)} = \begin{cases} 0, \text{ for } (I_{t+1(i,j)} - I_{t(i,j)}) <= T \\ 1, \text{ for } (I_{t+1(i,j)} - I_{t(i,j)}) > T \end{cases} \quad (1.1)$$

Presenter Tracking in a Classroom Environment

The major advantage of the temporal differencing method is that it does not rely on a constant lighting level nor on specific color characteristics of the user. However, one must be aware that the difference image may be essentially black if there is little or no movement between the two scenes.

Fortunately, temporal differencing is well suited for motorized camera tracking, as a response is required only when the subject moves substantially out of the center of the frame, also known as the region of interest. A stationary background is no longer required, since the effects of any change will only appear in the current difference image. Given this algorithm, the question now becomes how to exploit the data for robust tracking purposes with a pan and tilt camera.

Fig. 1. Scene of a user in a classroom setting.

In a classroom setting (fig. 1), overhead transparencies or video projectors pose a major threat to image processing algorithms as a result of the dynamic effects to light intensity and background characteristics. When a slide is changed, it may affect a large percentage of the resulting temporal difference image. However, if the presenter has not moved in this interval, no response from the system is desired. These requirements are satisfied by imposing a condition that camera adjustments are

made only when the difference image reflects movement over a number of consecutive frames; otherwise, the change is considered to be noise. This prevents the system from repositioning the camera erroneously when only the background has changed.

Once a difference image is created (see fig. 2), there are several methods of extracting the appropriate information to track the individual. We found histogram analysis suitable for this task. The vertical and horizontal histograms are constructed from the outliers of the difference image, and the locations of the maxima are used to depict the center of the region of movement.

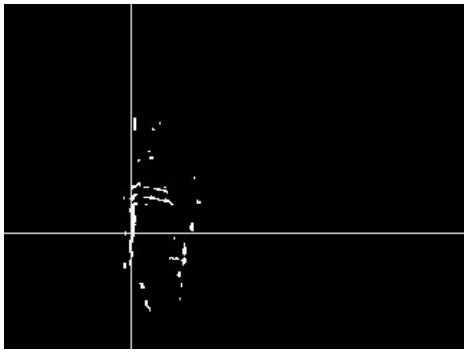

Fig. 2. Temporal difference image with crosshairs located at the maxima of the horizontal and vertical histograms.

If this center is outside the estimated position of the user by more than a certain threshold for a preset number of frames, the camera's orientation is adjusted appropriately. This ensures that undesirable camera responses do not occur as a result of small arm movements within the thresholded region of interest, denoted by the bounding box in fig. 3.

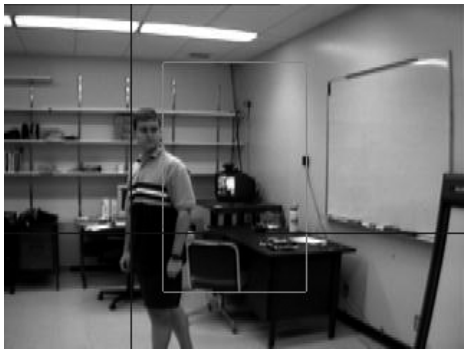

Fig. 3. Overlay of 'Region of Interest', and histogram determined crosshairs.

An adjustment made to the algorithm is in the proportion of height and width of the bounding box formed by the outliers in the difference image. Since the tracker is used predominantly while the presenter is standing or walking, an aspect ratio of 2:1 was found to be appropriate. If the bounding box does not adequately conform to this desired ratio, the camera position is not adjusted based on this information. This further helps the algorithm ignore changes that do not correspond to movements of the entire person. For example, if an audience member's head appears briefly in the lower part of the image, the camera will not respond erroneously.

Results

Preliminary experiments indicate that the algorithm is indeed sufficiently robust to follow an individual in a dynamic classroom environment and also demonstrates the potential to track multiple users. This could be achieved by noting not only the global maxima of the histograms, but also the local maxima. By doing this however, the possibility of overlapping users in a scene arises. If two pan and tilt cameras were to track two, separate users, the overall scheme becomes far more complicated.

There were many small adjustments needed to ensure robustness of the algorithm. For example, the camera is allowed to readjust only if the crosshair position appears outside the bounded region for a fixed number of scenes. If this number is too small, the camera readjusts overly frequently, producing an undesirable oscillatory effect. If this number is too large, a user will not be tracked while making a series of small movements across the scene.

In our initial experiments, the size of the bounding box was fixed. However, it is conceivable to change this size as a function of the zoom factor of the camera. The height and width were determined heuristically to account for users who move their arms far from their bodies, so that this would not result in oscillatory tracking effects. It is also worth noting that the camera panning (horizontal) motion worked quite well, while the tilt (vertical) occasionally moved too far. For example, movement of the presenter's legs lowered the maximum of the vertical histogram, hence tilting the camera downwards. This was solved by increasing the height of the region of interest, artificially. Another possible solution would have been to take the center of the bounding box formed from the outliers in the difference image.

Another consideration is that the camera must provide some form of feedback while in the process of a pan or tilt movement. Otherwise, the scene being viewed appears to change continuously until the camera motion is complete, resulting in a difference image that fills the entire scene. If such feedback is not available, it may be sufficient simply to wait a preset amount of time after each camera adjustment.

Conclusions

By investigating and comparing a variety of available techniques and tuning the algorithms as appropriate to our application, a final system has emerged that is not restricted to an idealized laboratory setting. This method performs well under a variety of conditions, and has demonstrated potential to be expanded for applications requiring zoom control and tracking of multiple users.

Acknowledgements

The authors would like to thank the Faculties of Engineering and Management of McGill University, who have provided space and funding for this research. Support has come from the Natural Sciences and Engineering Research Council, Fondes pour la Formation de Chercheurs et l'Aide a la Recherche (FCAR), Petro-Canada, and the Canadian Foundation for Innovation. This support is gratefully acknowledged.

References

1. A. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Campbell, Y. Ivanov, A. Schutte, A. Wilson. "The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment." M.I.T. Media Laboratory Perceptual Computing Section Technical Report. No. 398, November, 1996.

2. Carbonaro, A. and Zingaretti, P." Landmark matching in a varying environment." Proceedings Second EUROMICRO Workshop on Advanced Mobile Robots IEEE Comput. Soc., pp.147-53. Los Alamitos, CA, 1997.

3. T. Darrell, G. Gordon, J. Woodfill, and M. Harville. "Tracking People with integrated stereo, color and face detection". AAAI Symposium on Intelligent Environments, pp.44-50. Stanford, CA, 1998.

4. Davis, J. and Bobick, A. "The representation and recognition of human movement using temporal templates." Proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.928-34. Los Alamitos, CA, 1997.

5. S. Fels, Y. Sumi, T. Etani, N. Simonet, K. Kobayashi, and K. Mase. "Progress of C_MAP: A Context-Aware Mobile Assistant." AAAI Symposium on Intelligent Environments, pp.60-67. Stanford, CA, 1998.

6. Hunke, M. and Waibel, A. "Face Locating and Tracking for Human-Computer Interaction." 28th Asilomar Conference on Signals, Systems and Computers. Monterey, California. November 1994.

7. Kahn, R. and Swain, M. "Understanding people pointing: the Perseus system." Proceedings International Symposium on Computer Vision IEEE Comput. Soc. Press. pp.569-74. Los Alamitos, CA, 1995.

8. A. Medl, I. Marsic, M. Andre, C. Kulikowski, and J. Flanagan. "Multimodal User Interface for Mission Planning." AAAI Symposium on Intelligent Environments, pp.102-109. Stanford, CA, 1998.

9. Yang, Y. and Levine, M. "The Background Primal Sketch: An Approach for Tracking Moving Objects." Machine Vision and Applications, vol. 5, pp.17-34, 1992.