Arseneau, S. and Cooperstock, J.R. (1999), Real-Time Image Segmentation for Action Recognition. Proc. 1999 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM'99), Victoria, B.C., Canada, pp. 86-89.

Real-Time Image Segmentation for Action Recognition

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

In order to recognize actions of a human properly, an algorithm must be able to detect a person accurately in any given image. This data will be used to ascertain what action the user is performing. Many methods have been proposed that use a difference picture technique in which the image of the user is subtracted from a previously known image with the same background, without the user. These methods, however, are plagued with problems of noise and "ghosting," the undesirable introduction of image fragments as a result of changes in light intensity and moving objects in the background. The proposed algorithm combines several image processing methods in order to produce a clean difference image, while being far more robust to changes in light intensities and in the background scene.

Introduction

The next generation of computers will go far beyond the typical interfaces we use today. They will be controlled by methods that are more natural for a human than the cumbersome mouse and archaic keyboard. Recently there has been a push towards speech and gesture recognition as they both make maximum use of human capabilities of complex movements and communication skills. While speech recognition is rapidly maturing to a reasonable state, action recognition is still in its infancy. Plagued by longstanding problems of image processing such as proper segmentation and classification, it has been slow in its development. Some successful algorithms [3][7][9] have been developed for specific, highly controlled environments, but the challenge of creating such an algorithm that can operate robustly in an unstructured environment remains an open problem.

Image Segmentation Algorithm

The first challenge faced is that of proper segmentation. Many methods require that a user have a known, uniformly textured, or colored background (chroma-keying) or take in a single image of solely the background for background removal [2][5][7]. The latter introduces a number of problems related to lighting, noise and stability. If image processing is ever to grow beyond the lab, these problems must be addressed. One such solution is the *background primal sketch* [10], which allows for some of the inconsistencies that arise, such as lighting changes in the room. This primal sketch is constructed by taking the median value of the pixel color, as it is far more robust than the mean, over a series of images. The median, as well as a threshold value, are then used to construct the difference image. Let F^N_{jk} represent a sequence of N collected images, and (j,k) the pixel location. The resulting background primal sketch, denoted as B_{jk} is calculated as follows:

$$B_{jk} = \text{median}\ (F^1_{jk}\ ,\ F^2_{jk}\ ..\ F^N_{jk}\ ) \qquad (1)$$

The threshold is determined using a histogramming procedure based on the least median squares method [10]. The advantage of a pixel-by-pixel threshold approach over a uniform threshold method is that the areas that tend to change in intensity or position more than others, but are still encompassed in the background have a higher threshold, thus preventing much of the ghosting effects described earlier (see figure 1).

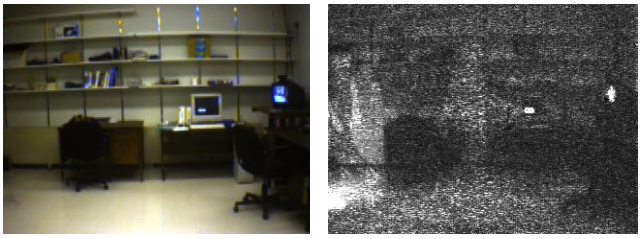


Figure 1 – (left) Background primal sketch constructed from 5 images. (right) Thresholded image

In the threshold image above, the brighter the pixel intensity, the higher the threshold value is for that location. While constructing both the background primal sketch and threshold image, the user appears briefly in the left-hand part of the scene, hence the higher threshold in that portion. One can also note that the threshold is higher for values representing the monitor, as is expected, since the frequency difference between monitor refresh and frame capture rates results in the appearance of a constantly sweeping bar down the screen. This change in the background would normally produce anomalies for a chroma-keying scheme, but the primal sketch method adapts for this problem.

It is also plausible to periodically update the background primal sketch over longer periods of time to account for a greater change in the background, such as the moving of a background

object. A ghosting effect would appear in the difference image only until at least half of the images taken to construct the primal sketch appear with the new location of the moved object. A suitable method could be to update the background for all of the pixels outside the bounding box scribed about the user, which will be explained further on in this paper.

Once the background primal sketch and threshold data are constructed, the difference image is taken from a scene with the user. (see figure 3) For the original image, see figure 2. The pixels that form the difference image are referred to as *outliers*, denoted as D_{jk} and obtained from calculating the difference between the incoming image, F^{i}_{jk} and the primal sketch B_{jk}, and comparing it against the threshold T_{jk} (see equation 2).

$$D_{jk} = 1, \text{ if } | F^{i}_{jk} - B_{jk} | > T_{jk} \qquad (2)$$
$$0, \text{ otherwise}$$

Another problem that arises often in image processing scenarios is that of tracking methods. Keeping in mind that the desired end result is to have information to interpret as actions, one must consider exactly how this is to be done. Color matching is very popular for tracking various parts of the body based on skin color [6]. Again, this is adequate for highly constrained environments, but many problems arise when the algorithm is applied to individuals of varying skin tones. Also, the tracking may be fooled into following the wrong parts of the body if only looking for skin tone, such as mistaking a hand for a face. Thus it seems to follow that a tracking algorithm may be more robust if it uses another characteristic of the image. It has been found that an edge-detected image (see, for example, figure 2) produces a more appropriate input for the determination of pose of a user. For example, the difference constructed from edge-detected images allows the locating of arm positions much more accurately, as in the case where the arm is located in front of the body (see figure 3).

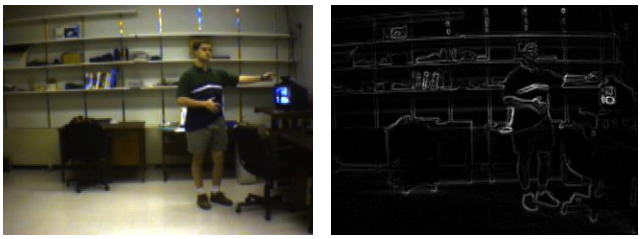


Figure 2 – (left) Original scene, (right) edge-detected scene using the Sobel Operator

The difference image as a result of a color-based primal sketch and input image does not give any clues as to what orientation the left-right arm is in, whereas the difference image formed from edge-detected images, reveals much more information about the pose of the user. An added bonus of performing an edge-detection operator on the images is that much of the noise is also eliminated in the difference image.

Noise Reduction Techniques

Once the difference picture made up of outliers is obtained, some noise reduction techniques are performed. First, isolated outliers are eliminated. This technique is quite effective as data points formed as a result of a person in the scene are highly concentrated.

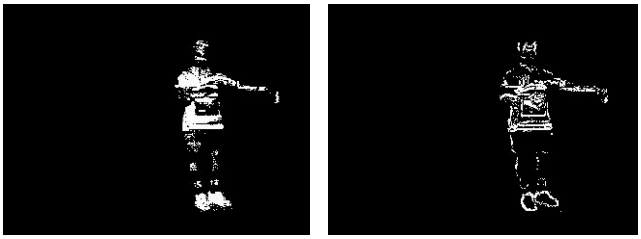


Figure 3 – (left) difference image from color-based method, (right) difference image from edge-detected method.

Therefore, by removing isolated outliers that have no other adjacent neighbours in the 8-connected neighborhood, much of the noise is reduced.

The next noise reduction technique makes use of Otsu's method of selecting a threshold via a histogram [8]. This approach employs the zeroth and first-order cumulative moments of the grayscale component of the remaining outliers. Using a histogram, a threshold is calculated, and those outliers that fall below the calculated threshold are eliminated (see figure 4). An interesting result of removing these particular pixels is that almost all of these outliers correspond to background or noise, while the overall shape of the person being tracked is maintained (see figure 5). This is to be expected as the pixel locations that change more dramatically are as a result of the moving person, while those few background, edge pixels that changed just enough to breach the threshold, typically fall short of Otsu's threshold.

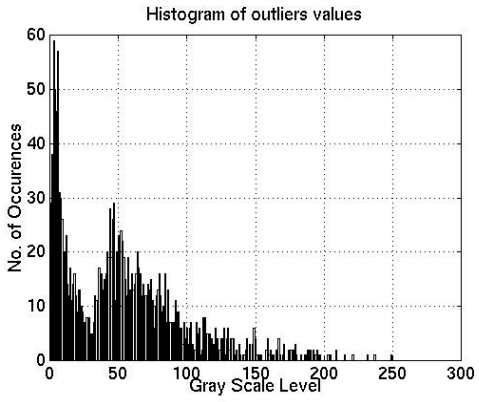
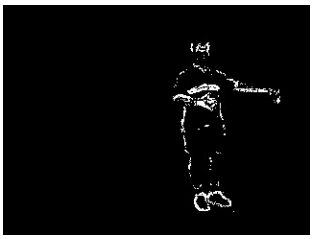


Figure 4 – Histogram of the grayscale values of all outliers. Otsu's threshold occurred at 33 in this example.

It should be noted as an aside that performing Otsu's technique to thresholding the red, green and blue components, often results in eliminating shadows of the person, again removing unwanted pixels.

By removing all outliers that appear below all three thresholds, much of the noise is dealt with. However, it also has the unwanted



effect of removing parts of the tracked person if the individual is wearing colors with dark components.

Figure 5 – Difference image after noise reduction techniques performed

Normally, the end result is a clear representation of where the person is located in the image. The next step is to determine the region of interest in the scene, namely the exact location and pose of the person. A preliminary step is to take the vertical and horizontal histogram maxima and use this as the center location of a bounded box. This cuts down on further processing that may be done to classify the action.

The next step is to determine the outline of the person within the scene. The variation of the active contour approach proposed by Levine and Yang [10] labels dilations of the outliers and then performs a controlled erosion to determine the optimal outline that encompasses the most of the original outliers. This is useful for filling in gaps between pixels but is quite costly to compute depending on the number of dilations performed.

The contour can also be extracted by "sweeping" a row of pixels from each side of the cropped image until it contacts the remaining outliers, the end result being the outline of the moving person. This pixel-sweeping approach is much quicker to calculate and provides a fairly accurate approximation of the desired outline. An example of pixel sweeping from the top of the image is shown in figure 6. As is shown, some discontinuities are obtained with this method. However several solutions are being investigated. For one, the pixels could be dilated once and then a sweep performed so as to more accurately depict the user's arms. There are also many simple smoothing techniques available to rectify some of the stray pixels. At this stage, using the color-based technique offers the advantage of having far more outliers within the person's outline, therefore making a sweep much more representative of the actual contour. However, as the final goal will be to classify the action, it may turn out that this particular step is unnecessary.

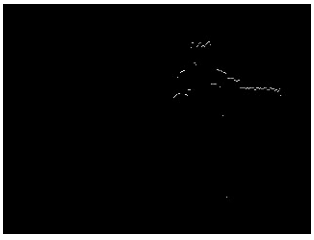


Figure 6 – Region of interest found, and pixel sweep performed from the top of the image

Classification of Actions

The next obstacle after segmentation has been performed is to classify those segments properly. Whether it is determined from shape, color, or size, a segmented image is of little worth unless useful information can be extracted from it. Many popular methods have been used for identifying objects, such as image template matching [1][2]. However, in order to identify a user's action requires the collection of information from a sequence of images. Furthermore, the challenge increases when one must decide the beginning and end of an action, for example, someone pointing towards the sky, or waving hello. Future work will most likely use features that are discretized directions and velocities of the hands. If after tracking the right hand, the classifier comes across a chain of events of a general tend of a pendulum motion**,** it might be categorized as a *waving* action.

Results and Conclusions

Formal, empirical comparisons of our approach with other algorithms remain to be completed. Qualitatively, the background primal sketch technique [10] combined with pixel sweeping has proven highly effective in determining the hand locations of a person being tracked. This image segmentation algorithm has so far proven to be robust for high frequency changes in the background, and we are confident that it can serve an important role in more complex action classification tasks.

References

1. Birchfield, S. "Elliptical Head Tracking Using Intensity Gradients and Color Histograms." *IEEE Conference on Computer Vision and Pattern Recognition*, 1998.

2. Davis, J. and Bobick, A. "The Representation and Recognition of Action Using Temporal Templates." *IEEE Conference on Computer Vision and Pattern Recognition*, 1997.

3. Davis, J. and Bobick, A. "Virtual PAT: A Virtual Personal Aerobics Trainer." *Workshop on Perceptual User Interfaces*, 1998.

4. Davis, J. and Bobick, A. "SIDESHOW: A Silhouette Based Interactive Dual-Screen Environment." *M.I.T. Media Lab Report*, No.457, 1998.

5. Huang, T., Blostein, S., Werkheiser A., McDonnell, M., and Lew, M. "Motion Detection and Estimation from Stereo Image Sequences." *IEEE Proceedings of the Workshop on Motion: Representation and Analysis*,

1986.

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. "Gesture Recognition Using the Perseus Architecture." *ISCV*, November 1995.

8. Otsu, N. "A Threshold Selection Method from Gray-Level Histograms." *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-9, no. 1, 1979

9. Pinhanez, C. and Bobick, A. "It/I: A Theater Play Featuring an Autonomous Computer Graphics Character." *M.I.T. Media Lab Report,* No. 455, 1998

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

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.