31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Automated Estimation of Elder Activity Levels from Anonymized Video Data

Nicholas Harvey, Zhongna Zhou, James M. Keller, *Fellow, IEEE*, Marilyn Rantz, and Zhihai He, *Senior Member*, *IEEE*

***Abstract*—Significant declines in quality of life for elders in assisted living communities are typically triggered by health events. Given the necessary information, such events can often be predicted, and thus, be avoided or reduced in severity. Statistics on activities of daily living and activity level over an extended period of time provide important data for functional assessment and health prediction. However, persistent activity monitoring and continuous collection of this type of data is extremely labor-intensive, time-consuming, and costly. In this work, we propose a method for automated estimation of activity levels based on silhouettes segmented from video data, and subsequent extraction of higher order information from the silhouettes. By building a regression model from this higher order information, our system can automatically estimate elder activity levels.**

I. Introduction

AUTOMATIC assessment and tracking of overall activity levels is a useful tool in the recognition of behavior changes, and thus potential health issues, in eldercare practice. Such automated estimation of activity levels can be accomplished by the application of intelligent algorithms to video data of elderly subjects. Specifically, by applying background subtraction techniques to video sequences, silhouettes of a subject can be segmented from the raw video. We have shown previously that silhouettes are accepted by elders as a privacy preserving methodology [1]. By modeling the activity level as a function of the behavior of the silhouettes, it is possible to construct a system capable of automatically estimating and tracking the activity level observed in a temporal sequence.

Foreground segmentation and motion tracking are a popular focus in computer vision research, due to their usefulness in automated surveillance system, among other applications. As a result, a multitude of different approaches to the problem have been developed. In [2], each image was transformed into series of texture descriptors, and background subtraction was performed based on statistical models of these texture descriptors, rather than the models of the raw intensity images. In [3] an independent component analysis (ICA) based technique is applied in order to separate foreground and background. The resulting method is robust, even over changing backgrounds, despite the lack of a background update step in the algorithm.

The methods of silhouette segmentation, human tracking, and activity analysis used as the basis for this work were first discussed in [4]. Many of the basic concepts of [4] are similar to those presented here; however in this work, we focus on calculating a quantitative measurement of activity level, rather than identifying and labeling various types of activity.

Our process begins with the configuration of a camera in the primary living area of the residence of the subject whose activity level is to be monitored. Video is then recorded of the daily activities of the subject. From this video data, a silhouette is extracted on each frame of the video, using a color based background subtraction method. The accuracy and precision of the silhouette segmentation process is assumed here to be relatively good. Each silhouette is then further analyzed in order to produce a set of features describing its position, size, and shape at the moment when that frame of video was recorded. Additionally, the rate of change in each of these silhouette descriptors is measured over a time window around the frame. Finally, the features of each silhouette in the entire sequence are combined and used as the inputs to an automated system, which estimates the activity level observed in the sequence, by using a model of activity level as a function of the measured features.

In order to gather ground truth on which to base a model, a group of health care students and faculty members from the University of Missouri-Columbia were recruited to provide assessments of activity levels in a preselected set of sample video sequences. Once the activity levels of the subjects in the sample video sequences were rated, the relationship between these ratings and the sequence features was used as the basis for our model.

Details on each step of this process can be found in the next section.

II. Method

A. Silhouette Segmentation

The decision to apply the silhouette segmentation technique to this problem was made for several reasons. First, there are, of course, inherent privacy concerns with the placement of a camera in a residence. By processing video data into silhouettes in real time, and deleting the original video sequences, these concerns are alleviated to some extent. Second, it is simply not feasible to store raw video

Manuscript received April 7, 2009. This work was supported in part by the National Institute of Health under Grant 1R21AG026412-01A2.

N. Harvey is with the University of Missouri-Columbia, Columbia, MO 65211 USA (phone: 314-307-4157; e-mail: nmhdh8@mizzou.edu).

Z. He, Z. Zhou, J. M. Keller are with the University of Missouri-Columbia, Columbia, MO 65211 USA (e-mails: hezhi@missouri.edu, zz3kb@mizzou.edu, kellerj@missouri.edu)

recordings of a subject over an extended period of time, due to the massive amount of disk space that such data would require. Since silhouettes are simply a binary (two color) image, they are easy to faithfully compress and store over the long term. Third, it is conceptually much simpler to model activity level based on features of silhouettes than it is to model it on raw video data. An example silhouette segmentation result is shown in Figure 1.

A detailed description of the process of silhouette segmentation is outside the scope of this paper. For a general review of background subtraction techniques, see [5]. Briefly, our process involves conversion of a full color image, in this case a single frame from a video sequence, to a binary image, where white pixels represent foreground (those parts of the image where the human subject is located), and black pixels represent background (everything else). In our method, a statistical model of the background is built, and pixels not matching the background model are labeled as foreground. Additional post-processing techniques are then applied, in order to eliminate shadows and other false positive foreground pixels [4].

It should be noted that the accuracy of the estimation of the activity level is highly dependent on the accuracy of the segmented silhouettes. Unfortunately, it was necessary here to make manual adjustments to the parameters of the silhouette segmentation algorithm in order to achieve acceptable results. This was the result of inherent limitations of the dataset, caused by a less than strictly controlled data capture process combined with difficulty in finding frames on which to base background models. Additional issues, such as changes in illumination, also caused problems with the silhouette segmentation. As a result of these problems, suitably accurate silhouettes could be segmented from only 44 of the original 50 sample video sequences. In the future, better control of the data capture process, as well as improvements to the silhouette segmentation method should help to eliminate such problems. In particular, methods of silhouette segmentation based on depth or disparity maps, which are robust over illumination changes, are currently an active area of research [6, 7].

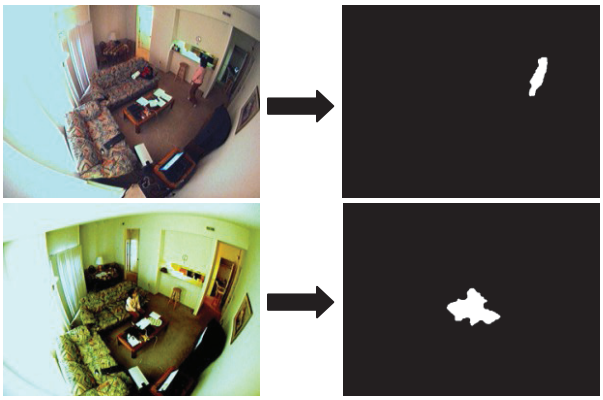

Fig. 1. An example of the silhouette segmentation process. The original image is on the left, and the result of silhouette segmentation is on the right. The top example is relatively accurate, while the one below is an instance of a less optimal result.

B. Silhouette Level Features

Once a silhouette is segmented, the next step is to calculate a set of features which describe the silhouette. We chose sixteen such features. The first four features are the median values of X and Y coordinates, the area of a bounding box, and the ratio of such a box's width to its height. The next four features are the means of all X and Y coordinates of pixels labeled as foreground (the centroid), the number of pixels labeled as foreground, and the angle of the primary axis of orientation of the silhouette. The remaining eight silhouette features were estimates of the rates of change in the first eight features. These features were calculated by finding the differences between the values of each of the first eight features in frame t-L and t+L, where t is the time step for which the features are being calculated and L is the width of the window over which the feature changes are being measured. Equation (1) displays this feature vector. Here, L is set to a value of 7, meaning that the width of the time window is approximately 2 seconds (the frame rate of the video capture is approximately 7 frames per sec).

Note that some of these features are conceptually very similar, and may contain redundant information. Because there was no way of predicting which features would prove to be useful, and because the dimensionality would later be reduced via principal component analysis, initially all of these features were calculated.

$$\overrightarrow{F(t)} = \begin{bmatrix} x_{median}(t) \\ y_{median}(t) \\ w(t) * h(t) \\ w(t)/h(t) \\ x_{mean}(t) \\ y_{mean}(t) \\ |sil(t)| \\ tan^{-1}\left(\frac{\xi_1(t)}{\xi_2(t)}\right) \\ |F_1(t-L) - F_1(t+L)| \\ |F_2(t-L) - F_2(t+L)| \\ |F_3(t-L) - F_3(t+L)| \\ |F_4(t-L) - F_4(t+L)| \\ |F_5(t-L) - F_5(t+L)| \\ |F_6(t-L) - F_6(t+L)| \\ |F_7(t-L) - F_7(t+L)| \\ |F_8(t-L) - F_8(t+L)| \end{bmatrix}$$

$$where: \overrightarrow{F(t)}: Feature\ vector\ at\ time\ t, \vec{\xi}: direction\ of\ greatest\ length\ of\ the\ silhouette, w: width\ of\ the\ bounding\ box, h: height\ of\ the\ bounding\ box, sil: set\ of\ foreground\ pixels$$

(1)

C. Sequence Level Features

After the silhouette features have been calculated for each frame, it is necessary to combine them into a set of features which describe the silhouette over a sequence of frames, i.e., a video clip. Both the scale and number of resulting sequence features should be independent of the length of the sequence. Several methods were used to derive such features. Minimum, maximum, mean, standard deviation,

and skewness values of each of the silhouette features over a sequence were calculated. Histograms of the values of each of the silhouette features were also calculated for all frames in a sequence, normalized so that the histograms approximated probability density functions (i.e. the sum of all bins was 1.0). Features describing the histograms were then derived, including the histogram centroid (the sum of the center values of each bin, weighted by the respective bin heights), the height of the peak of the histogram, the location of the peak of the histogram, the standard deviation of each silhouette feature histogram, the standard deviation of the histogram bin heights, and the skewness of the histograms. Again, some of these features may contain redundant information, but by including them we increase the amount of potential information available, and the dimensionality reduction step should eliminate problems with redundant information sources.

Because some of the silhouette features are measured on significantly different scales (e.g. the value of the area of the bounding box will usually be much larger than values for the position of the center) it was necessary to normalize the features, again using the standard method of subtracting the mean and dividing by the standard deviation. This helped to ensure that no features had a disproportionate effect on the results.

Once the feature values for this (relatively large) feature set were calculated, we attempted to find a smaller set of "useful" features. We applied principal component analysis (PCA) to the full feature set for all 44 data points, where each data point represents one video sequence. The data is transformed from its original space into a space spanned by the principal components. Thus we find a list of transformed features, ranked in order of decreasing contribution to the variance of the overall distribution. The features at the beginning of this list are considered to be the "best", or most useful features with regard to modeling activity level. By using the first N of these features as the basis for a model, we are using the N "best" features [8].

Using the selected features and corresponding activity ratings of experts, a basic linear regression model of activity level is constructed. This model can then be used to calculate an estimate of the activity level in other video sequences.

D. Expert Ratings

In order to construct an accurate model of activity level, it was necessary to gather a number of sample data on which to base such a model. Therefore, a group of health care students and faculty members from the University of Missouri-Columbia were recruited in order to generate sample data. These volunteers were chosen for their knowledge of the elderly subjects this project is ultimately meant to benefit, and their knowledge of this particular project and its goals. Ten volunteers were originally recruited, eight of whom gave complete responses. These eight assessments were used as the basis for all models.

These raters were asked to assess the activity levels in sample video clips in a simulated residential setting. Each clip was assigned a score on a scale of 0 to 10, with 0 corresponding to the lowest possible activity level, and 10 the highest. In order to eliminate inter-rater bias resulting from differing standards of what constitutes a high activity level, the scores assigned by each rater were normalized using the standard method of subtracting the mean score assigned by the rater, and dividing by the standard deviation of their ratings. The normalized ratings were then rescaled into the range 0.0 to 1.0, where the lowest rating overall was set to the value 0.0, and the highest was set to 1.0. The mean of all ratings was considered to be the true measure of activity level in the video sequence.

III. RESULTS

In order to evaluate the efficacy of our method, 11-fold cross validation was performed on the 44 samples in the data set. For each fold, a linear regression model was constructed using a different set of 40 data points, and the remaining 4 points were left as validation data. The model used the means of the activity level scores, assigned to each video sequence by the raters, as the responses, and modeled these responses as a function of the features produced by the previously discussed PCA dimensionality reduction process. In order to determine the optimal number of "best" features to use, the mean, over all 11 folds, of the mean squared error in the predicted activity level for the validation data was found for models based on the best N features, for values of N going from 1 to 30. Additionally, a comparison was performed between the estimation error of the models and the variance inherent in the expert activity level assessments. One would expect a higher estimation error in a sequence on which expert raters tended to disagree. We used the standard deviation in the expert assessments as a measure of disagreement, and counted the total number of model predictions of validation data, over all 11 folds, which were more than two of these standard deviations away from the mean of the expert assessments. These were labeled outlying model predictions.

TABLE I
CROSS VALIDATION ERROR IN ACTIVITY LEVEL ESTIMATION

Number of features	MSE (% of range)	Standard Deviation	Outlying model predictions
1	1.51%	±1.19%	7
2	1.55%	±1.17%	7
3	1.40%	±1.01%	7
4	1.31%	±1.05%	6
5	0.98%	±0.75%	4
6	0.95%	±0.76%	4
7	0.98%	±0.77%	4
8	0.99%	±0.73%	5
9	1.29%	±1.26%	7
10	1.58%	±1.64%	7
11	1.82%	±2.11%	7
12	1.76%	±1.95%	8
13	2.27%	±3.28%	8
14	2.61%	±3.79%	8
15	1.74%	±1.84%	6
16	3.49%	±4.70%	9

17	4.03%	±5.52%	9
18	3.96%	±5.28%	8
19	6.96%	±14.47%	10
20	9.30%	±16.79%	9
21	17.47%	±27.95%	11
22	24.61%	±46.44%	12
23	8.12%	±11.39%	10
24	9.18%	±13.36%	13
25	8.92%	±12.69%	10
26	9.16%	±12.97%	12
27	10.95%	±14.89%	13
28	7.95%	±9.54%	13
29	30.61%	±83.80%	19
30	39.87%	±99.74%	20

IV. Analysis

As can be seen from Table 1, the mean, over all folds, of the MSE in the validation data is minimized when using 6 features. Models based on between 5 and 8 features usually have comparable MSE for validation data. Using more than 8 features tends to result in over-fitting of the model to the training data used for regression, causing a loss of generality and increase in the validation error. Additionally, because of the nature of the linear regression method, the use of more features of the sample data in the regression model will invariably cause a decrease in the resubstitution error (i.e. the MSE in the prediction of the data used to construct the model). The number of features selected should be chosen such that overall error is minimized, while still maintaining the generality of the model. Using between 5 and 8 features meets these criteria.

Of note, the inclusion of interaction terms, and higher order terms in general, in the regression models did not appear to cause any significant decrease in the prediction error. This is most likely the result of the application of PCA. PCA is only capable of identifying linear relationships; therefore the use of PCA to perform dimensionality reduction will produce features which vary linearly with the responses.

Overall, the prediction error was relatively low, and for the most part fell within two standard deviations of the mean value of the expert assessments. These results are promising for our work in this area, and the feasibility of accurately assessing activity levels using an automated system.

V. Conclusion and Future Work

By segmenting the silhouettes from all frames of a video sequence, and analyzing the behavior patterns of the silhouette over the sequence, it is possible to automatically estimate the activity level of a person visible in the original video. As discussed earlier, this has the potential to be a very useful health maintenance tool for elders. This work demonstrates the feasibility of solving this problem.

Ultimately, the goal of this project is to facilitate the development of a system which can be placed in the residences of consenting elders, and automatically monitor and track their activity levels over time. There are several possible directions which we intend to explore as this research continues. Alternate methods of silhouette segmentation may yield more reliable results in that step. Different features and alternate methods of feature selection or dimensionality reduction may yield a more reliable model. The features and the method of dimensionality reduction used here were basic and intuitive. More complex and powerful methods of modeling the activity level may be able to more accurately mimic the assessments of human experts. The method of linear regression certainly possesses some limitations, which could possibly be overcome by the application of non-linear methods of modeling, such as fuzzy logic inference or kernelized support vector machine regression models.

Acknowledgment

We would like to thank the following students and faculty members of the University of Missouri-Columbia for their help with this work: B. Wakefield, C. Abbott, D. Oliver, J. Giger, G. Alexander, J. Krampe, C. Galambos, R. Koopman, and M. Aud.

References

[1] Demiris G., Parker Oliver D., Giger J., Skubic M., Rantz M., "Older adults' privacy considerations for vision based recognition methods of eldercare applications," Technology and Health Care (in Press).

[2] Dongxiang Zhou, Hong Zhang, Nilanjan Ray, "Texture Based Background Subtraction," in *Proc.2008 IEEE International Conference on Information and Automation,* 2008, Zhangjiajie, China, pp. 601-605.

[3] Du-Ming Tsai, Shia-Chih Lai, "Independent Component Analysis-Based Background Subtraction for Indoor Surveillance," *IEEE Transactions on Image Processing,* vol. 18, no. 1, pp. 158-167, Jan. 2009.

[4] Xi Chen, Zhihai He, Derek Anderson, James Keller, and Marjorie Skubic, "Activity Analysis, Summarization, and Visualization for Indoor Human Activity Monitoring, " *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1489-1498, Nov. 2008.

[5] M. Piccardi, "Background subtraction techniques: A review," presented at the IEEE SMC 2004 International Conference on Systems, Man, and Cybernetics, The Hague, The Netherlands, Oct. 2004.

[6] Hansung Kim, Dong Bo Min, Shinwoo Choi, and Kwanghoon Sohn, "Real-time disparity estimation using foreground segmentation for stereo sequences," Opt. Eng. 45, 037402 (2006).

[7] Muñoz-Salinas, R., Aguirre, E., and García-Silvente, M. "People detection and tracking using stereo vision and color," *Image Vision and Computing,* 25, 6 (Jun. 2007), 995-1007.

[8] Jolliffe, I.T., *Principal Component Analysis*, Springer-Verlag, New-York, 1986.