Published in final edited form as:
Neurocomputing. 2011 June 1; 74(12-13): 2184–2192. doi:10.1016/j.neucom.2011.02.014.

Physical Activity Recognition Based on Motion in Images Acquired by a Wearable Camera

Hong Zhang[a,*], **Lu Li**[a,b], **Wenyan Jia**[b], **John D. Fernstrom**[c], **Robert J. Sclabassi**[b], **Zhi-Hong Mao**[d], and **Mingui Sun**[b,d,*]

[a] Image Processing Center, Beihang University, Beijing 100191, China

[b] Department of Neurosurgery, University of Pittsburgh, PA 15213, USA

[c] Department of Psychiatry, University of Pittsburgh, PA 15261, USA

[d] Department of Electrical and Computer Engineering, University of Pittsburgh, PA 15261, USA

Abstract

A new technique to extract and evaluate physical activity patterns from image sequences captured by a wearable camera is presented in this paper. Unlike standard activity recognition schemes, the video data captured by our device do not include the wearer him/herself. The physical activity of the wearer, such as walking or exercising, is analyzed indirectly through the camera motion extracted from the acquired video frames. Two key tasks, pixel correspondence identification and motion feature extraction, are studied to recognize activity patterns. We utilize a multiscale approach to identify pixel correspondences. When compared with the existing methods such as the Good Features detector and the Speed-up Robust Feature (SURF) detector, our technique is more accurate and computationally efficient. Once the pixel correspondences are determined which define representative motion vectors, we build a set of activity pattern features based on motion statistics in each frame. Finally, the physical activity of the person wearing a camera is determined according to the global motion distribution in the video. Our algorithms are tested using different machine learning techniques such as the K-Nearest Neighbor (KNN), Naive Bayesian and Support Vector Machine (SVM). The results show that many types of physical activities can be recognized from field acquired real-world video. Our results also indicate that, with a design of specific motion features in the input vectors, different classifiers can be used successfully with similar performances.

Keywords

Activity recognition; classification; feature extraction; feature matching; motion histogram; multiscale

1. Introduction

Video based activity recognition has been an active field of research in computer vision and multimedia systems [1-9]. Although numerous algorithms have been developed, a

*Corresponding author. dmrzhang@buaa.edu.cn drsun@pitt.edu.

fundamental requirement has been that the target object engaging in a certain activity must appear in the video, which is not achievable in many practical cases where the object never appears in the video because the camera can only be mounted on the target object itself. Examples of such objects include a spaceship, an aircraft, a submarine, a vehicle, a robot, an animal, or a person. For example, if the goal is to study an individual's physical activity over an entire day in a free-living environment, it is unrealistic to track the person with video cameras. Alternatively, with today's technological advancement, a subject can comfortably wear a small camera for the entire day. Although he/she does not appear in the recorded video, physical activity can be recognized indirectly by observing the recorded background scene. In general, a specific activity will result in a specific motion of the camera since it is mounted on the human body and the background scene will change accordingly when the camera is moved.

We have been investigating the use of a wearable video device to monitor food intake continuously in obese individuals [10]. However, modification of diet (energy input) represents only half of the energy balance equation of the human body. The other half is physical activity (energy output). A worn device that unobtrusively and automatically records physical activity will provide a powerful tool for the development of individualized obesity treatment programs that help people lose weight and keep it off.

Wearable sensors that objectively measure body motion and dynamics have been developed [11-14]. One common approach is to use accelerometers attached at multiple locations of the body to measure both acceleration and orientation [11, 12]. The accuracy of physical activity recognition by accelerometer-based systems generally improves as the number of accelerometers increases. However, the obtrusiveness of such systems makes it inconvenient to wear and use in daily life. We have thus developed a new wearable device, which contains a video camera and other sensors, to monitor both food intake and physical activity (see Fig. 1). The device is mounted in front of the chest using a pin, a lanyard or a pair of magnets, allowing it to measure the trunk motion of the upper body. While the general design of the device and its food intake measurement function are published elsewhere [10, 15, 16], this paper describes the algorithms utilized to process the acquired video data and recognize several common types of physical motion and activity.

Numerous vision algorithms are available for activity recognition using features extracted from the observed target directly[2, 3]. Unfortunately, these algorithms do not apply to our case where the target (a person) does not appear in the video. The key problem is therefore to find descriptions of the physical activity in the recorded video without direct observation of the target. These descriptions can be obtained if the following two assumptions are satisfied: 1) the motion profiles of the activities to be recognized differ from each other, and 2) the background scene is rich enough so that sufficient image features can be extracted reliably.

We approach the indirect activity recognition problem by investigating camera motion and developing an activity detection scheme based on 2D image features. Considering that the camera in the wearable device is usually not controlled intentionally and, as a result, the acquired images are often blurry, we match correspondent points between adjacent frames using multiscale local features. In order to reduce errors in pixel-pair matching, we impose uniqueness and epipolar constraints which eliminate ambiguous pixel pairs. After the correspondence selection process, a motion histogram is defined according to motion vectors obtained from the selected pixel pairs. For each activity video, an accumulation of motion vectors is evaluated based on a set of motion histograms to obtain global motion characteristics which finally lead to physical activity recognition.

This paper is organized as follows. Section 2 provides an overview of our algorithms. Detailed descriptions of these algorithms are presented in Section 3. Experimental results are provided in Section 4. In Sections 5 and 6, we summarize our work and discuss future directions on physical activity recognition using the wearable camera approach.

2. Outline

Our framework for recognizing physical activity is shown in Fig. 2. It consists of three data processing steps: feature extraction, motion representation and activity recognition. In the first step, local image features are extracted from which a set of "salient key points" are determined. In the second step, we match these key points between neighboring frames. The matched points define a set of motion vectors. A motion histogram is then defined according to these vectors. In the last step, the cumulative motion over all frame pairs is evaluated. Physical activity is recognized based on motion histograms and global motion characteristics.

3. Methods

3.1. Feature Extraction

Obtaining reliable correspondences of features is essential in our activity recognition system because inaccurate correspondences produce ambiguous motion estimation. We use local image features which are widely investigated [17-26]. We prefer local features to global features because the local ones can be detected and represented more easily. The key problem here is to find salient points in each image. Shi and Tomasi [22] described a method called "Good Features", which computes the minimum eigenvalue of the covariance matrix instead of the cost function defined in the Harris detector [23]; Lowe [17] presented a Scale Invariant Feature Transform (SIFT) method using scale space analysis. This method is invariant to scale, orientation and affine distortion [18-20]. Bay *et al*. [24, 25] proposed a Speed-Up Robust Features (SURF) method using the 2D Haar wavelet.

Although the existing methods have been well studied for activity recognition from directly recorded images as the input, these methods usually require these images to have reasonably high quality. In our case, however, the wearable device is uncontrolled and thus the images acquired are often blurred. We have found that the blur of our images resulted from many factors, and it is hence difficult for us to choose the most suitable model for de-blurring. Occasionally, an incorrect model even aggravates noise. Hence we use a multiscale detector to capture motion features in an "overlooking" scale in which the extracted information is less affected by blurring than that in the raw image.

In our application, the wearable camera is often used continuously to record data for more than ten hours a day. Existing multiscale methods such as SIFT [17] and SURF [24] are not sufficiently efficient in our case. We have therefore developed a new multiscale feature detection approach, which works rapidly and reliably on data containing noise and distortion, such as affine distortion and illumination variation.

We first define *interest pixels* in a single image I from the covariance matrix C at pixel *p*:

$$\mathbf{C}=\begin{bmatrix} \left(\sum_{N_p} I_x\right)^2 & \sum_{N_p} I_x I_y \\ \sum_{N_p} I_x I_y & \left(\sum_{N_p} I_y\right)^2 \end{bmatrix} \quad (1)$$

where I_x *and* I_y are, respectively, the intensity partial derivatives along the x- and y-axes, and N_p is the 3x3 neighborhood of the feature pixel *p*. We use the minimum eigenvalue to determine the *interest pixels* in the input image and reduce the effect of affine distortion [22]. Given λ_p of the minimum eigenvalue at pixel *p*, the most interested pixels (features) are chosen where $\lambda_p > c \cdot \lambda_{max}$, with λ_{max} being the largest eigenvalue for the pixels in I, and *c* being set to 0.01 in our experiments.

Next, we combine the above feature detection approach with a multiscale analysis. For discrete images, our multiscale representation is expressed as a Gaussian pyramid according to the scale space theory [18, 27]. The Gaussian pyramid is a set of images $\{g_0, g_1, \cdots, g_k\}$ constructed from the original image g_0, where index *k*, $k = 0, 1, 2, \ldots, N$, stands for the level in the pyramid. The adjacent elements in the pyramid are related by

$$g_{k+1}(u,v) = \sum_m \sum_n w(m,n)\, g_k(2u+m, 2v+n) \tag{2}$$

where *w(m, n)* is a Gaussian weighting function in the neighborhood of g_k and *N* refers to the number of levels in the pyramids. We find all features in g_k for all levels of *k* using the method in [22]. In order to normalize the results, we remap all pixels (*u, v*) of interest at scale level *k* back to the scale of original image, at $\left(\frac{u}{k}, \frac{v}{k}\right)$. The selected pixels of interest (feature pixels) are shown in Fig. 3 where the red and green dots represent, respectively, the interest pixels in the original image and in the immediate level of the Gaussian pyramid.

3.2. Motion Representation

In our case, the motion of the wearable camera is closely related to the physical activity that the wearer performs. We first find pixel correspondences between adjacent frames in which an epipolar constraint is imposed. Next, we define a motion histogram obtained from pixel correspondences in each pair of video frames. Finally, the motion histograms of all pairs of video frames are used to form feature vectors for activity recognition.

3.2.1. Pixel Correspondence—For a feature pixel *I(x, y)* in the first frame and its corresponding pixel $I'(x', y')$ in the second frame, we impose a uniqueness constraint on $I'(x', y')$, given by

$$\begin{aligned} I'(x',y') = \arg \min_{I'(x',y') \in \Omega'} \; & D\left(I(x,y), I'(x',y')\right) \\ \text{s.t.} \quad D\left(I(x,y), I'(x',y')\right) < \alpha \cdot D & \left(I(x,y), I'(x'',y'')\right), \\ I'(x'',y'') \in \Omega' \setminus & I'(x',y'), \; 0<\alpha<1 \end{aligned} \tag{3}$$

where Ω' is a neighborhood in I' centered at $I'(x', y')$, $\Omega' \setminus I'(x', y')$ denotes region Ω' excluding pixel $I'(x', y')$, and α is an empirically selected threshold to enhance reliability of the matched features.

In addition to Eq. (3), we impose the following epipolar constraint for corresponding pixels *I(x, y)* and $I'(x', y')$ [28]:

$$\left(P'_H\right)^T \mathbf{F} P_H = 0 \tag{4}$$

where $P_H^{'}$ and P_H are, respectively, the homogeneous coordinates in the forms $P_H = (x, y, 1)^T$ for $I(x, y)$ and $P_H^{'}=\left(x^{'}, y^{'}, 1\right)^T$ for $I^{'}(x^{'}, y^{'})$, and **F** is a 3×3 fundamental matrix between $I^{'}$ and I. We use the Random Sample Consensus (RANSAC) algorithm [29, 30] to eliminate the pixel pairs which do not satisfy Eq. (3). At least eight pairs of point correspondences are needed to estimate the fundamental matrix and remove outliers [31]. The green dots and lines in Fig. 4 show an example of outlier elimination.

3.2.2. Motion Distribution in Frame Pairs—Once the points of correspondence are determined from video frame pairs, we characterize the physical activity in the video according to the direction and magnitude of motion vectors. Given a pixel correspondence $[I(x, y), I^{'}(x^{'}, y^{'})]$ in adjacent frames, the motion vector of this pair of pixels is defined as:

$$\begin{aligned} M_i &= \left(x_i^{'} - x_i, y_i^{'} - y_i\right) \\ \theta_i &= \begin{cases} \arctan\left(\frac{y_i^{'}-y_i}{x_i^{'}-x_i}\right), x_i^{'} - x_i>0, y_i^{'} - y_i \geq 0 \\ \pi+\arctan\left(\frac{y_i^{'}-y_i}{x_i^{'}-x_i}\right), x_i^{'} - x_i<0 \\ 2\pi+\arctan\left(\frac{y_i^{'}-y_i}{x_i^{'}-x_i}\right), x_i^{'} - x_i>0, y_i^{'} - y_i<0 \\ \frac{\pi}{2}, x_i^{'} - x_i=0, y_i^{'} - y_i>0 \\ \frac{3\pi}{2}, x_i^{'} - x_i=0, y_i^{'} - y_i<0 \\ undefined, x_i^{'} - x_i=0, y_i^{'} - y_i=0 \end{cases} \\ |M_i| &= \sqrt{(x_i^{'} - x_i)^2+(y_i^{'} - y_i)^2} \end{aligned} \tag{5}$$

where θ_i is the direction of vector M_i in the range of $[0, 2\pi)$, and $|M_i|$ is the magnitude of M_i.

Because our wearable camera records the scene of the background rather than the activity performer, the magnitude of motion may vary significantly since it is strongly related to the distance between the background scene and the performer. We therefore propose the use of an orientation based motion histogram to characterize motion between neighboring frames. We define an n-bin histogram **h** as follows. First, θ_i is equally divided into n bins with each bin covering an angular range of $2\pi/n$. Thus, n specifies the resolution of motion orientation. Next, within each bin, the number of motion vectors in direction θ_i, which belongs to $[s \cdot 2\pi/n, (s + 1) \cdot 2\pi/n)$, $s = 0, 1, \ldots, n - 1$, is counted, where s is the index of histogram bin. In order to reduce measurement error, we require $|M_i|$ to be no less than a threshold t, measured in pixels. Finally, we normalize the motion histogram **h**, represented as an n-dimensional vector **m** by:

$$\mathbf{m}=\frac{\mathbf{h}}{\sum_{s=1}^{n} h_s} \tag{6}$$

where h_s is the s-th element in vector **h**.

3.3. Activity Recognition

In addition to the motion histogram which provides a characterization of motion in a pair of frames, we need an effective representation of motion for the entire video consisting of numerous frames. There exists an approach using three-dimensional spatial-temporal features for action recognition [3, 5]. Although this 3D approach is effective, it has a high computational complexity. In our application, the daily life physical activities, such as walking and exercising, usually consist of short segments of simple, repeated actions. Therefore, we define an activity recognition vector representing the statistical characteristics

of the input video segment. Our activity recognition vector consists of two parts: a set of summed sample values (here sample values refer to the values in the histogram bins) and the standard deviation of these values. The summed sample values are utilized for two purposes: 1) to reflect the average of the motion histogram across the segment of the activity video; and 2) to smooth out noise. The use of the standard deviation is to capture the range of sample distribution. A collection of activity recognition vectors is used to train a pattern classifier.

3.3.1. Activity Recognition Vector—Let a video contain f frames. For normalized motion histogram vector $\mathbf{m_j}$ in each frame, we compute mean μ_j and standard deviation σ_j by

$$\begin{aligned}\mu_j &= \frac{\sum_s m_{js}}{n}, s=1,2,\ldots,n, j=0,1,\ldots,f-1\\ \sigma_j &= \sqrt{\frac{1}{n}\sum_s\left(m_{js}-\mu_j\right)^2}, s=1,2,\ldots,n, j=0,1,\ldots,f-1\end{aligned} \tag{7}$$

where m_{js} is the s-th element of vector $\mathbf{m_j}$ and j is the frame number. Under the assumption of ergodicity, we may calculate combined statistics by

$$\begin{aligned}\mu &= \frac{1}{f-1}\sum_j \mu_j, j=1,\ldots,f-1\\ \sigma &= \sqrt{\frac{1}{f-1}\sum_j\left(\sigma_j^2+\mu_j^2\right)-\mu^2}, j=1,\ldots,f-1\\ \mathbf{d} &= \left[\sum_j \mathbf{m_j}, \sigma\right], j=0,1\ldots f-1\end{aligned} \tag{8}$$

where μ and σ are the combined mean and standard deviation, and the $n+1$ dimensional vector $\mathbf{d}$ is defined as the activity recognition vector (ARV).

3.3.2. Training and Classification—In this section, each ARV is computationally classified into a certain physical activity. In the classifier, the inputs and output, respectively, are the ARVs and the activity types. We use the Support Vector Machine (SVM) as the classifier because it has been successfully applied to visual activity recognition tasks with proven efficacy [3, 32]. Other types of classifiers may also be utilized. In our SVM implementation, an automatic parameter selection scheme was utilized based on the K-fold cross validation procedure[33]. In this procedure, all SVM parameters were tested starting from the minimum and then incremented exponentially till the maximum or a pre-defined stopping criterion (maximum number of iterations) was reached. The parameters with the best cross validation result were chosen for the SVM classifier.

4. Experimental Results

The first step of our experiments was the extraction of stable key points as candidates for finding point correspondences. After the correspondences were determined, motion vectors were obtained and mapped to histogram bins according to magnitudes and orientations. Activities were then characterized and recognized by statistically combining information in motion histograms. To test the stability and compatibility of our method, we chose three pixel features and three classifiers for feature detection and activity recognition. We tested six video sets containing real-world physical activities, including sitting-up, sitting still, walking, bowing, crouching and waist exercise. The training and testing data were acquired using three background scenes for each type of activity. The classifiers were trained and tested using the cross validation scheme. All the six physical activities were performed by a single human subject. We did not use multiple subjects because we believed that the motion

profiles among different healthy subjects of similar body heights were similar. Since the background of the scene exerted much stronger impact on the recognition accuracy than subjects, we tested multiple background scenes (See Fig. 5 for example). Each set of data consisted of 100 video segments. The video lengths for different activities were different. However, within each activity, the video length was identical. We collected video data at a rate of 10 frames per second with a screen size of 320×240.

4.1. Experimental Procedure

Our comparative experiments were arranged as follows. First, we tested three local image descriptors, including those by Shi and Tomasi [22], the Speed-up Robust Feature descriptor [24] and our own multiscale descriptor. For the purpose of comparison, we used the SVM classifier and two histogram configurations with resolutions $n = 8$ and 10. Next, we varied the number of orientations n for the chosen classifier and descriptors and observed the trend in error rate. Finally, we tested the performance of different classifiers with varying ARV dimensions.

4.2. Performance of Image Features

Local features played an important role in our wearable camera case. Efficiency and accuracy were both required because the amount of data to be processed was large. We first tested two state-of-the-art feature extraction methods by Shi and Tomasi [22] and by Bay *et al.*[24].

The Good Features (GF) method [22] enables the detection of occlusions, disocclusions and features that do not correspond to a real-world point. The only parameters to be set are the minimal distance between two key points and the minimum quality level of features to be accepted. In our experiments, we fixed these parameters at 100 and 0.01, respectively. The window size used to estimate the covariance matrix was 3×3.

The SURF detector [24] is a widely used, well performed local feature detector that provides scale and rotation invariability. This detector uses a similar feature generation procedure to that used by the SIFT descriptor [17, 18] except an approximated Hessian matrix is utilized to simplify the calculation of features. The SURF descriptor gives similar or better results under the change of view point, scale and luminance, and the processing speed is higher[24]. In our experiments, the default 64-dimensional descriptors were used, the Hessian threshold was set to 100, and the number of octaves and the number of layers in each octave were chosen to be 3 and 4, respectively.

We first tested both methods with chosen parameters on six types of activities. The training sample size, test sample size and video length in each category are listed in Table 1. The columns are the six individual activities.

Table 2 shows the sample recognition rates of GF and SURF methods with respect to the number of histogram bins (8, 10 and 16) and the type of physical activity. The last column provides overall results. It can be observed that the SURF method over-performed the GF method. We believe that our videos acquired by the wearable camera contributed to the performance difference. These videos were often blurred because of the free movement of the human body. With this type of input, the GF method was often ineffective in capturing feature points. In contrast, the SURF detector was less disturbed because it was multiscale-based allowing the capture of blurred points more effectively.

We evaluated the processing times of both methods for feature extraction using 600 videos. The results varied considerably. The average processing times (in seconds) are shown in Table 3 for 8 and 10 histogram bins. Our evaluation was performed on a PC platform of

2.13GHz CPU and 4GB memory. We found that, although the SURF method was 8.3% more accurate than the GF method on average (from Table 2), the SURF method required an average of 2.5 times more computational time. This is a critical problem since we must be able to process real-world images acquired continuously for as long as several days.

With clear advantages and drawbacks of both methods demonstrated, we tested our own multiresolution Good Feature detection method (MRGF). The number of resolution levels was set to 3. The results are shown in Table 2 and Table 3.

It can be observed that, although the average recognition rate of the SURF method (91.3%) was slightly better than the MRGF method (90.5%), the MRGF method reduced the processing time by about one-half. It can also be observed, from the last column of Table 3, that the multiscale methods (MRGF and SURF) outperformed the single scale method (GF) in all experiments. The recognition precision increased significantly from GF to MRGF in the condition of low resolution where n equals to 8 and 10.

4.3. Resolution of Motion Orientation

The motion histogram requires a choice of the number of histogram bins n which specifies the resolution of motion orientation n. We experimentally investigated the trend of recognition performance with respect to the choice of n. A comparison of the detection rates under different histogram resolution, image feature types and activities is given in Table 2, where the numbers of histogram bins tested were 8, 10 and 16. It can be observed that, for all types of image pixel features, the recognition rate increased with the number of histogram bins over all activities. Although for specific activities and image pixel features, the recognition rate fluctuated, the overall tendency was clear according to the data acquired so far. In order to provide more proof on the relationship between recognition rate and resolution of motion orientation, we conducted another experiment focusing more on the number of histogram bins. The threshold of magnitude t defined in Section 3.2.2 was fixed at 3 in all experiments.

Fig. 6 gives a detailed representation of the recognition performance with respect to the resolution of motion orientation. Two major conclusions can be drawn from our results. First, the average detection rate increased with n from 4 to 18. The GF, SURF and MRGF methods all achieved higher recognition rates with a larger n. The lowest rate was with $n = 4$ for all pixel features. Second, when the resolution reached a certain level, the improvement in accuracy became small. It was shown that the SURF and MRGF features were more stable. Since pixel feature representation is essential in the overall performance of recognition, the MRGF method, which provides such representation at a low computational cost, is more advantageous than the other two methods.

4.4. Classification of Recognition Vectors

Specific physical activities were recognized based on activity recognition vectors (ARVs). Our main purpose of experiments was to validate the effectiveness of ARVs. Three independent classification strategies, including Naive Bayes, K Nearest Neighbor and Support Vector Machine were used in our experiments.

Naive Bayes [34]—The Naive Bayes classifier requires a small amount of training data to estimate the means and variances necessary for classification. By assuming that the variables in feature vectors are independent, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

K-Nearest Neighbors (KNN) [34]—The K-Nearest Neighbor classifier is robust to noise. In our experiment, we set the number of nearest neighbors *K* to 10 empirically. It was found that the size of the training set could be chosen in a wide range between 6 and 20, and the choice was not sensitive to *K*.

Support Vector Machine (SVM) [35-37]

We found that, for different data sets, the SVM testing results varied considerably. We solved this problem by using a K-fold cross validation technique for parameter selection[33]. In this technique, each training set was divided into 10 subsets (folds). The parameters with the best cross validation result were used in the classifier.

The recognition results of the Naive Bayes, KNN and SVM classifiers are shown in Fig. 7. It can be observed that the SVM classifier outperformed other two classifiers when the histogram resolution was large. However, the Naive Bayes and KNN classifiers excelled in speed because the SVM classifier spent considerable time on choosing parameters. All classifiers achieved recognition rates above 85% for the six types of activities. The best performances for the KNN and Naive Bayes classifiers occurred when *n* was 10, while the best performance for the SVM classifier occurred when *n* was 14. Despite these differences, our results show that all three classifiers were applicable to our indirectly recorded activity data.

5. Discussion

In summary, our indirect activity recognition method mainly consisted of three data processing steps: feature detection, motion representation and activity classification. We characterized the activity patterns in the video using pixel features because this type of features was easier to obtain than other types of features, such as lines and blobs. Although pixel features can be detected in different ways, including the use of GF and SURF, there is a major trade-off between the accuracy and computational complexity. In our case, we used the MRGF detector because it provided a balance between these two factors and the features provided by this detector can be used to construct the motion histograms. The only parameter required by the MRGF detector was the resolution of motion orientation, which was found to be an insensitive parameter as long as it was sufficiently large.

In the activity recognition part of our study we embedded statistical information into motion vectors and used the activity recognition vector to classify and recognize activities. In our case, the recorded activity usually consisted of repeated activities of short times. By summing the motion vectors, temporal variations were reduced effectively and the features became more distinct in the activity recognition vector because of the repetition.

In general, a specific activity will result in a specific motion of the camera since it is mounted on the human body. When the camera is moved, the background scene will change accordingly. We believe that this change contains two components: a component reflecting the case-dependent scene of the physical world observed by the camera, and a component specific to the activity being performed. The change of background scenes can be represented by extracting image features and obtaining motion vectors. Although we have not intentionally separated the two components, our results show that activity can be recognized based on the extracted image features, under the assumptions stated below.

Firstly, it is assumed that the motion profiles of the classes of activities are sufficiently different. The activities investigated by this manuscript satisfy this assumption and thus can be recognized using the presented method. However, there exist activities which do not satisfy this assumption and the motion profiles alone do not lead to effective classification.

In these cases, recognition has to be achieved by adding image content analysis which is not covered by this paper.

Secondly, we assume that the background scene is rich enough so that sufficient image features can be extracted robustly. It is clear that feature extraction and matching will fail completely when the background scene is textureless (e.g., a smooth, satin white wall). Fortunately, this case is rare in practice and the stated background scene assumption is satisfied in most cases.

We mounted our device containing a camera in front of the chest using a pin, a lanyard or a pair of magnets. However, in our opinion, this is not the only choice and there is no vantage point on the human body to mount the device. In our system design, we had to consider several practicality constraints in selecting the mounting point. We chose the chest because: 1) it is a stable location that facilitates the acquisition of high-quality images reflecting the trunk motion of the upper body; and 2) the location is related high in the body to keep camera view clear from being blocked by arms/hands and other objects, so the occlusion problem, which is significant in many current activity detection systems, is greatly reduced.

As a final remark, for the study of obesity, it is desirable to obtain a quantitative measure of energy expenditure for each type of physical activity performed by the subject who is wearing the device. This can be easily calculated using the identified activity and the length of time engaged in the activity (both determined by the wearable device) together with a table of energy expenditure values associated with each activity [38].

6. Conclusion

In this paper, we have presented a new computational tool to study physical activity by analyzing video data acquired using a wearable camera. Our investigation has focused on the feasibility and the framework to perform pattern recognition by inferring from the images showing only the surrounding scene. We have presented a multiscale scheme to identify pixels of interest and showed that this scheme can be applied efficiently and robustly. We have also proposed the use of statistical properties of motion vectors. As shown in our experimental results, the activity recognition vectors that we utilized can be effectively classified with a high accuracy. The classification error decreases as the resolution of motion orientation vectors increases. Our experimental results have also shown that different classifiers can be applied to our activity recognition vectors with comparable error rates. The methods proposed in this work are useful in real-world applications where the camera can only be mounted on a target object and the data are acquired under imperfect conditions.

Acknowledgments

This work was supported by National Institutes of Health Grant No. U01 HL91736 of the United States, the National Natural Science Foundation of China Grant No. 60872079 and National Key Laboratory on Optical Feature of Environment and Target Foundation of China No. 9140C6105080C61.

Biographies

Hong Zhang received the B.S. degree, M.S. degree and Ph.D. degree in electrical engineering from Hebei University of Technology, China, Harbin University of Science and Technology, China and Beijing Institute of Technology, China in 1988, 1993 and 2002, respectively. She is currently a Professor of the School of Astronautics, Beihang University, Beijing, China. She was in the Department of Neurosurgery at the University of Pittsburgh as a visiting scholar from 2007 to 2008. Her research interests include activity recognition, image restoration, image indexing, object detection and stereo vision.

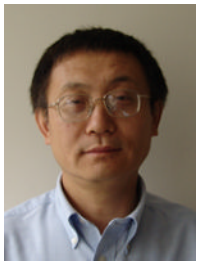
Lu Li received the B.S. degree in computer science from Hebei University of Technology, China in 2006. He is currently a Ph.D. candidate in Image Research Center, School of Astronautics at Beihang University, China. He studied in the Department of Neurosurgery at the University of Pittsburgh as a visiting student from 2009 to 2010. His research interests include activity recognition, image understanding and stereo vision.

Wenyan Jia received the B.S. degree and M.S. degree in biomedical engineering from Capital Medical University, Beijing, China, in 1998 and 2001, respectively, and the Ph.D. degree in biomedical engineering from Tsinghua University, Beijing, China, in 2005. She is currently a research assistant professor of the Department of Neurosurgery at the University of Pittsburgh, PA. Her research interests include image processing, pattern recognition, and biomedical signal processing.

John D. Fernstrom received the S.B. (Life Sciences) and Ph.D. (Nutritional Biochemistry & Metabolism) from M.I.T., Cambridge MA. He was a post-doctoral fellow at the Roche Institute of Molecular Biology, Nutley NJ, and a faculty member at M.I.T prior to becoming Professor of Psychiatry and Pharmacology at the University of Pittsburgh, and Research Director of the UPMC Weight Management Center. He has authored over 200 articles and reviews in his areas of expertise, and has edited the proceedings of five scientific conferences.

Robert J. Sclabassi received the B.S.E. degree in electrical engineering from Loyola University, Los Angeles, CA, the M.S.E.E. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, and the M.D. degree in medicine from the University of Pittsburgh, Pittsburgh, PA. He was with the Advanced Systems Laboratory at TRW, Los Angeles, and a Postdoctoral Fellow at the Brain Research Institute, University of California, Los Angeles (UCLA), where he has also been a Faculty Member in the Department of Neurology and Biomathematics. He is currently the CEO of Computational Diagnostics, Inc. and a Professor Emeritus of neurological surgery, psychiatry, electrical engineering, mechanical engineering, and behavioral neuroscience at the University of Pittsburgh. He has authored or coauthored more than 400 papers, chapters, and conference proceedings. Prof. Sclabassi is a Registered Professional Engineer.

Zhi-Hong Mao is an Assistant Professor in the Department of Electrical and Computer Engineering and Department of Bioengineering at the University of Pittsburgh, Pittsburgh, PA. He received his dual Bachelor's degrees in automatic control and mathematics (1995) and M.Eng. degree in intelligent control (1998) from Tsinghua University, Beijing, China, and S.M. degree in aeronautics and astronautics (2000) and Ph.D. degree in electrical and medical engineering (2005) from Massachusetts Institute of Technology, Cambridge. He was a recipient of the National Science Foundation CAREER Award and Andrew P. Sage Best Transactions Paper Award of the IEEE Systems, Man and Cybernetics Society in 2010.



Mingui Sun received the B.S. degree from Shenyang Chemical Engineering Institute, Shenyang, China, in 1982, and the M.S. and Ph.D. degrees in electrical engineering from the University of Pittsburgh, Pittsburgh, PA, in 1986 and 1989, respectively. In 1991, he joined the faculty of University of Pittsburgh, where he is currently a Professor of neurosurgery, electrical and computer engineering, and bioengineering. His current research interests include advanced biomedical electronic devices, biomedical signal and image processing, sensors and transducers, biomedical instruments, brain-computer interface, electro-neurophysiology, implantable devices, radio-frequency systems for biomedical applications, electronic and data processing systems for diet and physical activity assessment, and wearable computers. He has authored or coauthored more than 300 papers.

References

1. Poppe R. A survey on vision-based human action recognition. Image and Vision Computing. 2010; 28(6):976–990.
2. Laptev I, Caputo B, Schüldt C, Lindeberg T. Local velocity-adapted motion events for spatio-temporal recognition. Comput. Vis. Image Underst. 2007; 108(3):207–229.
3. Sun J, Wu X, Yan S, Cheong LF, Chua T-S, Li J. Hierarchical spatio-temporal context modeling for action recognition. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. 2009:2004–2011.
4. Wu, J.; Osuntogun, A.; Choudhury, T.; Philipose, M.; Rehg, JM. Proceedings of the International Conference on Computer Vision (ICCV). Rio de: 2007. A scalable approach to activity recognition based on object use.
5. Laptev I, Lindeberg T. Local descriptors for spatio-temporal recognition. First International Workshop on Spatial Coherence for Visual Motion Analysis. 2004
6. Niebles JC, Wang H, Fei-fei L. Unsupervised learning of human action categories using spatial-temporal words. Proc. BMVC. 2006
7. Xu D, Chang S-F. Video event recognition using kernel methods with multi-level temporal alignment. IEEE Trans. on Pattern Analysis and Machine Intelligence. 2008; 30(11):1985–1997.
8. Xu D, Chang S-F. Visual event recognition in news video using kernel methods with multi-level temporal alignment. IEEE Int. Conf. on Computer Vision and Pattern Recognition. 2007
9. Duan L, Xu D, Tsang I, Luo J. Visual event recognition in videos by learning from web data. IEEE Int. Conf. on Computer Vision and Pattern Recognition. 2010
10. Sun M, Fernstrom J, Jia W, Hackworth S, Yao N, Li Y, et al. A wearable electronic system for objective dietary assessment. Journal of the American Dietetic Association. 2010; 110:45–47. [PubMed: 20102825]
11. Intille SS, Bao L, Bao L. Physical activity recognition from acceleration data under seminaturalistic conditions. Tech. rep. 2003
12. Tapia EM, et al. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart monitor. Proc. Int. Symp. on Wearable Comp. 2007
13. Li, L.; Zhang, H.; Jia, W.; Nie, J.; Zhang, W.; Sun, M. Automatic video analysis and motion estimation for physical activity classification; IEEE 36th Annual Northeast Bioengineering Conference; 2010.
14. Zhang H, Li L, Jia W, Fernstrom JD, Sclabassi RJ, Sun M. Recognizing physical activity from ego-motion of a camera. IEEE International Conf. of EMBS. 2010

15. Sun, M.; Yao, N.; Hackworth, SA.; Yang, J.; Fernstrom, JD.; Fernstrom, MH.; Sclabassi, RJ. Proc. Int. Symp. Digital Life Technologies: Human-Centric Smart Living Technology; Tainan, Taiwan: 2009. A human-centric smart system assisting people in healthy diet and active living.
16. Li, Y.; Zhang, H.; Hackworth, S.; Li, C.; Yue, Y.; Sun, M. The design and realization of a wearable embedded device for dietary and physical activity monitoring; Proc. 3rd ISSCAA; Harbin, China. 2010.
17. Lowe, D. Object recognition from local scale-invariant features; Seventh International Conference on Computer Vision (ICCV'99); 1999. p. 1150-1157.
18. Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision. 2004; 60:91–110.
19. Lindeberg T. Feature detection with automatic scale selection. International Journal of Computer Vision. 1998; 30:79–116.
20. Lindeberg T, Bretzner L. Real-time scale selection in hybrid multi-scale representations. Proc. Scale-Space'03, Springer Lecture Notes in Computer Science 2695. 2003:148–163.
21. Ke Y, Sukthankar R. Pca-sift: A more distinctive representation for local image descriptors. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. 2004:506–513.
22. Shi J, Tomasi C. Good features to track. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. 1994:593–600.
23. Harris, C.; Stephens, M. A combined corner and edge detection; Proceedings of The Fourth Alvey Vision Conference; 1988. p. 147-151.
24. Bay H, Tuytelaars T, Gool LV. Surf: Speeded up robust features. In ECCV. 2006:404–417.
25. Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-up robust features (surf). Comput. Vis. Image Underst. 2008; 110(3):346–359.
26. Zhang H, Mu Y, You Y, Li J. Multi-scale sparse feature point correspondence by graph cuts. Science in China Series F:Information Sciences. 2010; 53(6):1224–1232.
27. Burt PJ, Edward, Adelson EH. The laplacian pyramid as a compact image code. IEEE Transactions on Communications. 1983; 31:532–540.
28. Hartley, R.; Zisserman, A. Multiple View Geometry in Computer Vision. Cambridge University Press; 2003.
29. Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM. 1981; 24(6):381–395.
30. Torr PHS, Murray DW. The development and comparison of robust methods for estimating the fundamental matrix. International Journal of Computer Vision. 1997; 24:271–300.
31. Ma, Y.; Soatto, S.; Kosecka, J.; Sastry, SS. An Invitation to 3-D Vision: From Images to Geometric Models. SpringerVerlag; 2003.
32. Laptev I, Marszalek M, Schmid C, Rozenfeld B, Rennes I, Grenoble II, Ljk L. Learning realistic human actions from movies. CVPR. 2008
33. Intel. Opencv library version 2.0, Tech. rep., Intel. 2009. http://sourceforge.net/projects/opencvlibrary/
34. Fukunaga, K. Introduction to statistical pattern recognition. 2nd ed.. Academic Press Professional, Inc.; 1990.
35. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995:273–297.
36. Drucker H, Burges CJC, Kaufman L, Smola AJ, Vapnik V. Support vector regression machines. NIPS'96. 1996:155–161.
37. Meyer D, Leisch F, Hornik K. The support vector machine under test. Neurocomputing. 2003; 55(1-2):169–186.
38. Ainsworth B, Haskell W, Whitt M, Irwin M, Swartz A, Strath S, et al. Compendium of physical activities: an update of activity codes and met intensities. Medicine & Science in Sports & Exercise. 2000; 32(9 Suppl):S498–504. [PubMed: 10993420]

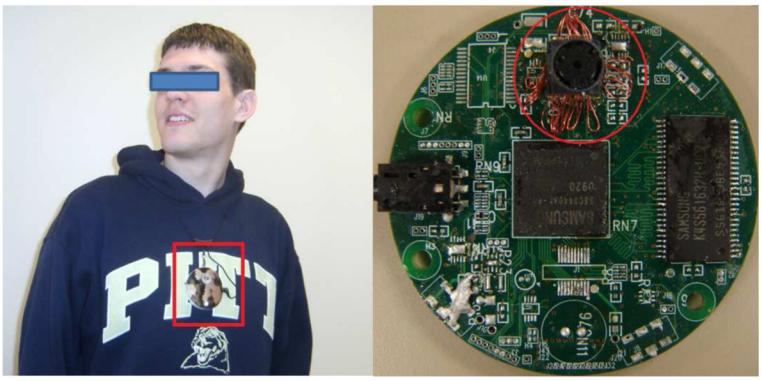


Figure 1.
Wearable camera developed in our laboratory. Left: person wearing the device. Right: Circuit board of the device where the camera is positioned inside the red circle.

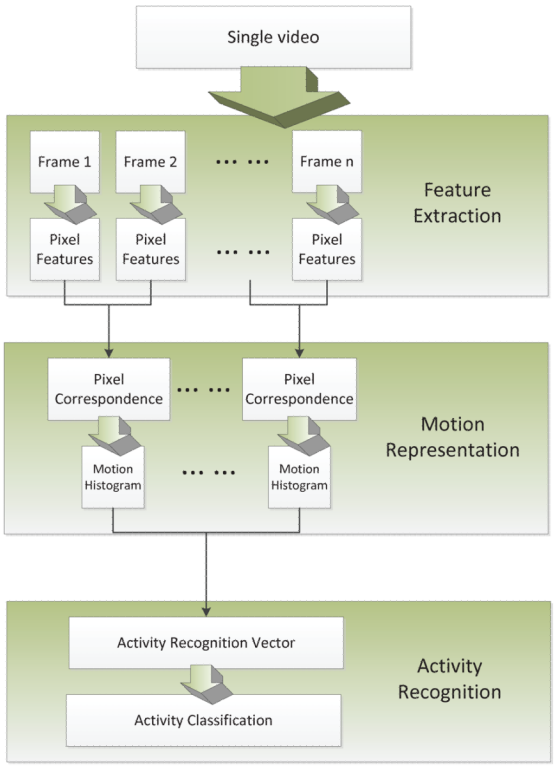
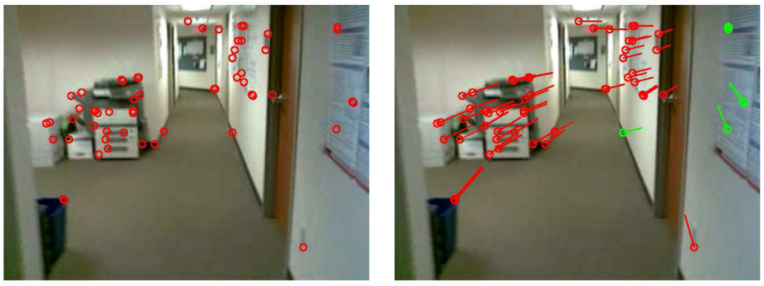


Figure 2.
Framework for activity recognition

Figure 3.
Feature pixels detected by ”Good Features” [22] in the original image (red dots) and the next level of the Gaussian pyramid (green dots).



(a) (b)

Figure 4.
(a) Pixel features are shown in red dots (b) Motion vectors (dot: location; short line: direction) computed from feature pixels. Red and green vectors indicate, respectively, the vectors satisfy and do not satisfy Eq. (4).

Figure 5.
Test scenes for walking.(a) Corridor, (b) Room (far), (c) Room (near), (d) Outdoor

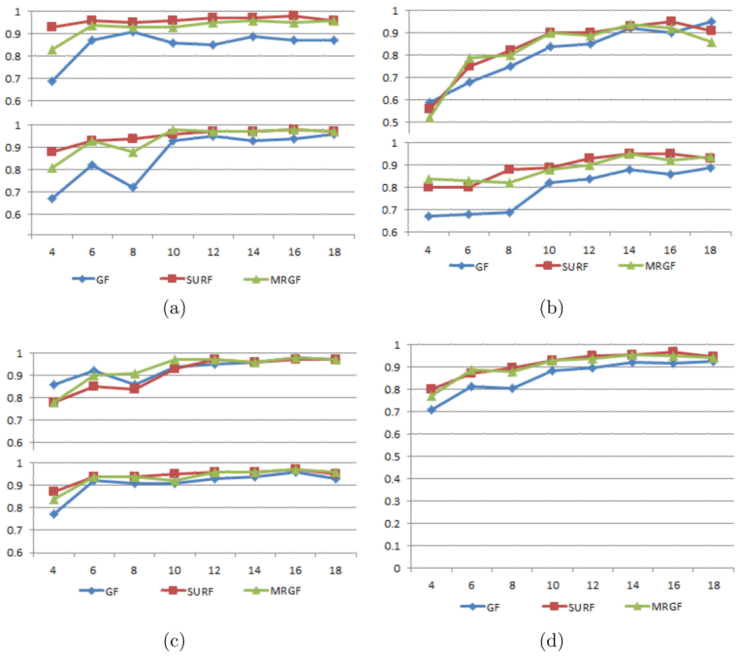


Figure 6.

Recognition rate (vertical axis) of six activities using the SVM method with respect to histogram resolution (horizontal axis).(a) top three curves: sitting-up, bottom three curves: sitting still; (b) top three curves: walking, bottom three curves: bowing;(c) top three curves: crouching, bottom three curves: waist exercise; (d) Overall rate (average over all six activities).

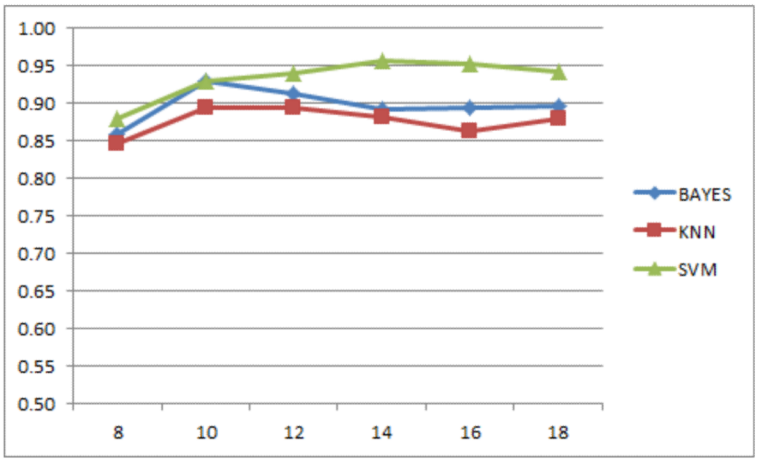


Figure 7.
Recognition rate using Naive Bayes, K Nearest Neighbors and Support Vector Machine. *Horizontal axis*: resolution of motion orientation. *Vertical axis*: recognition rate. Image features: MRGF.

Table 1

Training sample size, test sample size and video length in six categories of activities. Columns from left: Sitting-up(SU), sitting-still(SS), walking(WK), bowing(BW), crouching(CR) and waist exercise(WE). Video length is measured in frames.

	SU	**SS**	**WK**	**BW**	**CR**	**WE**
Training Sample	100	100	100	100	100	100
Test Sample	100	100	100	100	100	100
Video Length	51	51	51	61	81	121

Table 2

Recognition rates of GF, SURF and MRGF methods. Numbers of histogram bins n = 8, 10 and 16 were tested. Columns from left to right: Sitting-up(SU), sitting-still(SS), walking(WK), bowing(BW), crouching(CR), waist exercise(WE) and total average(ALL).

		SU	SS	WK	BW	CR	WE	ALL
GF	n = 8	91%	72%	75%	70%	86%	91%	81%
	n = 10	86%	93%	84%	82%	94%	91%	88%
	n = 16	87%	94%	90%	86%	98%	96%	92%
SURF	n = 8	95%	94%	82%	88%	84%	94%	90%
	n = 10	96%	96%	90%	89%	93%	95%	93%
	n = 16	98%	98%	95%	95%	97%	97%	97%
MRGF	n = 8	93%	88%	80%	82%	91%	94%	88%
	n = 10	93%	98%	90%	88%	97%	92%	93%
	n = 16	95%	98%	92%	92%	98%	97%	95%

Table 3

Average processing times (in seconds) for local image feature extraction. Six activities were tested (top row, containing sitting-up(SU), sitting-still(SS), walking(WK), bowing(BW), crouching(CR) and waist exercise(WE)). Each activity contained 100 videos. The numbers of histogram bins were $n = 8$ and $n = 10$

		SU	SS	WK	BW	CR	WE
GF	$n = 8$	114	145	129	181	224	293
	$n = 10$	118	134	121	174	225	288
SURF	$n = 8$	290	482	673	696	723	796
	$n = 10$	281	456	717	771	741	779
MRGF	$n = 8$	182	197	267	264	316	384
	$n = 10$	175	216	257	278	329	395