12-2016

Preliminary Investigation of Walking Motion Using a Combination of Image and Signal Processing

Bradley Schneider
Wright State University - Main Campus, schneider.163@wright.edu

Tanvi Banerjee
Wright State University - Main Campus, tanvi.banerjee@wright.edu

Preliminary Investigation of Walking Motion Using a Combination of Image and Signal Processing

Bradley Schneider, Tanvi Banerjee
Wright State University
Department of Computer Science and Engineering
Fairborn, OH, USA
{schneider.163, tanvi.banerjee}@wright.edu

Abstract— **We present the results of analyzing gait motion in first-person video taken from a commercially available wearable camera embedded in a pair of glasses. The video is analyzed with three different computer vision methods to extract motion vectors from different gait sequences from four individuals for comparison against a manually annotated ground truth dataset. Using a combination of signal processing and computer vision techniques, gait features are extracted to identify the walking pace of the individual wearing the camera as well as validated using the ground truth dataset. Our preliminary results indicate that the extraction of activity from the video in a controlled setting shows strong promise of being utilized in different activity monitoring applications such as in the eldercare environment, as well as for monitoring chronic healthcare conditions.**

Keywords – video analysis, motion and tracking algorithms and applications, gait analysis, activity detection.

I. INTRODUCTION

As technology has become more accessible, affordable, and capable, it has in many ways also become integrated into the environment around us and embedded in nearly every aspect of daily life. One such area in which technology has had an increasing presence is in assisted living and elderly care. Specifically, being able to ubiquitously monitor Activities of Daily Living (ADL) has great application in care for the elderly and disabled and has been researched in great length. Applications of activity detection can range in from providing automated assessments of rehabilitation progress to in-home monitoring of patients with diseases such as Alzheimer's. By detecting and tracking ADLs, subtle patterns of changes in activity may be examined, providing more individual information to healthcare providers to make well-informed decisions. Specifically, through the use of sensors such as cameras [9], [10] through continuous monitoring of individuals over time can help identify deteriorating health conditions in a timely manner which is crucial for successful interventions by clinicians [15].

A dominant challenge in this domain is in finding a method of monitoring which does not create too large of a burden on the patient being assessed while addressing the privacy concerns of the user. Activity detection through video analysis has been an active field of research due to the non-invasiveness of the approach. Methods based on other types of sensors typically require the subject to be instrumented with cumbersome gadgets such as accelerometers that may hinder the normal behavior pattern of users. In video-based techniques such as [9], [10], the camera is stationary, and usually the environment is instrumented instead of the subject, making for a more practical alternative. However, since the environment must be instrumented, these methods are constrained to operate within that closed space or the camera's field of view, and cannot be used elsewhere without the installations of additional cameras.

To remove this limitation on video-based approaches, we investigate a first-person video approach using a wearable camera. The video is recorded from the perspective of the monitored subject, meaning that no setup of the environment is necessary. Since the video sensor is embedded in a pair of glasses (that can be prescription), not only does it blend into the environment but also helps mitigate the need for instrumentation and calibration of the setup. Subjects are able to move and perform activities naturally, without interference from the sensor. In this study, we analyze the collected video with multiple motion extraction techniques to determine their effectiveness and the plausibility of detecting motion through the first-person video sensor.

For this initial investigation, the scope of activity detection has been limited to a single activity – walking, a typical ADL performed in the course of daily life. Specifically, we explore the feasibility of the wearable camera to distinguish between different gait speeds across four participants. Using computer vision techniques combined with signal processing methods, we extract features and compare it against curated ground truth data from the participants. Simply being able to detect gait information through first person video can allow for identification of movement patterns in elderly patients or gait analysis for patients undergoing rehabilitation. In the fitness domain, it may provide a simple means of tracking the number of steps taken by the subject. Outside of the healthcare realm, this technology could have applicability in military, emergency response, and law enforcement areas as well. The rest of the paper is organized as follows: Section II discusses the related work on activity recognition, Section III discusses the data collection process using the Pivothead wearable camera, Section IV discusses the computer vision techniques used to extract video frame motion, and Section V discusses the signal processing techniques used to draw conclusions about gait from those calculated motion vectors. Section VI describes the next steps in our research using the Pivothead.

II. Related Work

Much work has been completed on detecting activities of daily living. Many of these existing works deal with interactions of individuals with objects and surfaces in the environment to detect when specific activities are being performed [9], such as hand washing or cooking. The techniques focus largely on the context of the action, rather than the specific way in which the action is occurring. Because of this, they are often limited to detect actions in a very specific environment. In addition to environmental limitations, many proposed activity detecting systems require a large amount of instrumentation of the subject, requiring the wearing of accelerometers, respiratory sensors, and even rucksacks [14].

Existing efforts have shown that it is possible to detect such activities through first-person video captured via wearable devices [1, 4]. At the core of these approaches is the detection of specific objects and the user's hands, and their spatial relationship. Object detection plays a critical role in these findings.

Other efforts have focused on analyzing motion in first person video to define activities. Detecting a variety of activities complicates the task. For example, short and long-term activity detection with a single method can be difficult, though hybrid methods have been shown to detect activities with a fair amount of accuracy [3].

Limiting the detected activity to walking (and variations of walking) has also been investigated. One approach has used a downward-facing camera, capturing the movement of the subject's legs, to estimate gait information [13]. This method has the great benefit of naturally filtering out any other motion in the scene since only the legs, feet, and ground are in the frame. However, this mounting position for the camera is still less convenient than a forward-facing camera. It is acknowledged that the same methods should be useful with a forward-facing camera [13].

III. Data Collection

A. Hardware - Pivothead

Our attempt is to design a system with a realistic sensor that can feasibly be worn without any inconvenience to the user. For this reason, we use a video camera embedded within a regular pair of glasses to collect the first-person video segments. Multiple commercial solutions currently exist or are in development. We chose to use the Pivothead SMART Architect Edition glasses for our video capture [2]. They offer a convenient form factor and a pluggable platform to extend battery life and enable live-streaming of video over Wi-Fi. Conveniences such as these contribute to the goal of non-invasive activity detection.

Fig. 1. Pivothead Glasses with Embedded Camera

B. Computer Vision Method

Video clips were initially collected in a controlled environment in order to limit any motion within the video frame that was caused by an external source. The environment consisted of a treadmill in the center of a room, facing a pair of doors. By collecting the video indoors, movement from wind, trees, cars, or other objects was eliminated. Additionally, the setup of the treadmill directly in front of the doors eliminated any chance of motion from other people in the area.

Data were collected from four different subjects (2 male and two female participants of age 25 to 53) at two different speeds – 2.3 mph and 3.9 mph. Each subject wore the glasses, set the treadmill to a constant speed, and captured a video clip for approximately 10 seconds of walking. Subjects 1 and 2 are of similar height, and were 5-6 inches taller than subjects 3 and 4, who are also of similar height. This is important to note, since the taller subjects are expected to take less steps when walking at the same speed as the shorter subjects. Therefore, we expect that subjects 1 and 2 will have a lower frequency walking cycle than subjects 3 and 4 at the same walking speed.

In order to support potential applications in which live-streaming of the video is desirable, the video was streamed over Wi-Fi to a nearby device where it was recorded. Because the video was being streamed, it was collected at a resolution of 848 x 480 at 30 frames per second, rather than in full high definition.

C. Analysis

After collecting the video samples, they were manually annotated to generate a set of vectors indicating the exact true motion of the subject between each frame. For each frame of video that was analyzed, the location of a well-defined, clearly

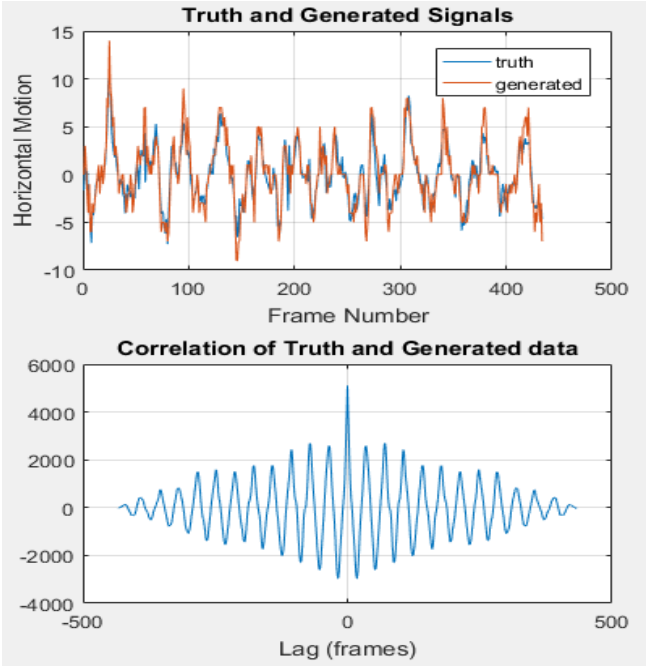

Fig. 2. Overlaid Truth and Generated Motion Data and Correlation Measure of the Two Signals

visible feature was manually annotated. Similarly, the location of the same feature was recorded for the next frame. Then, the difference of the two locations was recorded as a two-dimensional vector, indicating the amount of horizontal and vertical movement, in pixels that occurred between frames.

For a walking activity, we expect the motion of the wearable camera to be a periodic, cyclic wave. Looking to biological mechanism of the human gait motion, we confirm this expectation [12]. As the subject alternates between stepping with the left and right foot, the horizontal motion is expected to be back and forth. Similarly as either foot is picked up and moved forward for a step, the motion of the subject's head should be a cyclic up and down motion. The waves indicating horizontal and vertical movement are expected to be slightly out of phase with one another. This pattern is also confirmed by related work done using different types of sensors [17].

The manually collected truth data generally shows the pattern that was expected. Fig. 2 shows a graphical depiction of the horizontal and vertical movement over time for one set of collected truth data. While not exemplifying a perfect periodic wave, the data definitely depicts the back-and-forth swaying motion that provides the gait signature of an individual. The extra bounce in the vertical data is likely due to the camera being embedded in a pair of glasses. As each foot hits the ground during a step, there is vertical bounce allowed by the earpiece of the glasses. The shape of the earpiece is much less likely to allow such bounce in the horizontal direction, resulting in a smoother curve.

IV. Motion Extraction Techniques

The collected video sample was analyzed with multiple techniques in order to automatically extract the motion information. We form a comparison with the truth data and find which method is most effective at matching the data collected during manual analysis. Three general techniques were chosen for comparison due to their use in similar efforts: dense optical flow, sparse optical flow, and feature matching. Each method was implemented using the OpenCV C++ computer vision software library.

A. Dense Optical Flow

Dense optical flow is an algorithm that takes two consecutive frames of video as input and provides as output a motion vector for each single pixel in the frame. Though computationally expensive, it can achieve high accuracy since every image pixel is considered. The algorithm is described by Gunnar Farneback and is based on calculating the displacement estimation for each pixel neighborhood, which has been approximated by a polynomial expansion [6]. This method performs best with a slowly varying displacement field i.e. small local movements in the scene.

Fig. 3 shows an example of a frame overlaid with the vectors calculated from the dense optical flow algorithm. The vectors indicate the direction of movement of each pixel, with their length magnified for illustration. The detection of movement is limited to the areas of the image which are not solid surfaces, demonstrating the limitation of the algorithm in the regions that are lacking in texture or variation.

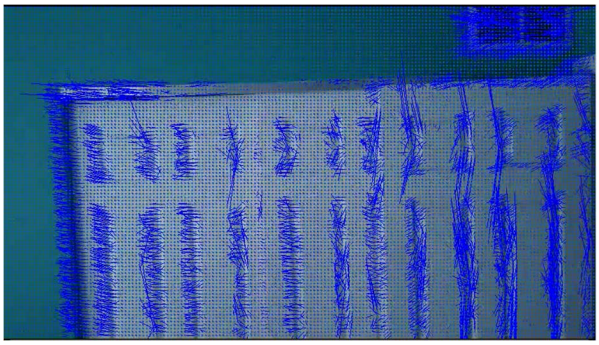

Fig. 3. Dense Optical Flow for a video sequence where the participant was walking while wearing the PivotHead glasses, and overlooking a door frame. The flow vectors are shown in blue.

B. Sparse Optical Flow

Sparse optical flow is another algorithm which provides an estimation of motion between two frames of video. Instead of extracting a motion vector for each pixel in the frame, motion is estimated for a set of key features from the image. This requires a two-step process of (1) feature extraction and (2) optical flow calculation.

The feature extraction method that was used was the Scale-Invariant Feature Transform [7], which extracts scale-invariant features from the video frame. This was chosen as a robust feature extractor despite the fact that the scale in the video is unchanging, since the subject and environment are both stationary.

The extracted features are passed into a sparse optical flow implementation derived from the work of Lucas and Kanade [5]. The Lucas-Kanade method requires the precondition that the time increment (and therefore, the movement) between frames is very small, and the image contains objects with smoothly changing intensity values. We can assume that these are met due to the frame rate of the video, and the tendency of natural scenes to have a smooth intensity gradient.

The sparse optical flow process is computationally much cheaper than dense optical flow, since calculations are only made on features that are significant to the image, greatly reducing the number of computations.

C. SURF Feature Matching

Speeded Up Feature matching (SURF) is an alternative to optical flow algorithms. Rather than estimating a motion vector for a point in every image in the video frame, SURF attempts to extract features in one frame and then extract those same features in the next frame [8]. By calculating the difference in location within the two frames for those two features, an estimated motion vector can be derived.

D. Evaluation of Motion Extraction Techniques

In addition to a set of manually constructed truth vectors for frame motion, a set of vectors was automatically extracted using each of the three discussed algorithms. For each frame, each algorithm creates several vectors – one per feature, or in the case of dense optical flow, one per pixel. To reduce the numerous vectors per frame into a single result indicating the overall frame motion, four aggregation methods were investigated: minimum, median, mean, and maximum vectors

by magnitude. We choose several methods to account for any tendencies in the algorithms to either under- or over-estimate the actual motion.

For each algorithm and each aggregation method, the sum of squared errors (SSE) was calculated against the truth data (see Table 1). The relative ordering of the SSE results indicates the best match of generated data with the truth data, starting with the lowest SSE.

In each of the aggregation methods, dense optical flow provided sub-optimal results when compared with sparse optical flow. This is likely due to the nature of the field of view in the controlled environment. As noted earlier, dense optical flow fails to return an accurate flow vector for pixels that are in a smooth, non-textured region of the input image. In the case of our sample, the solid areas of the wall provide insufficient input to the algorithm. It is unable to distinguish movement in those regions of the image. By discarding such regions, the performance of dense optical flow would likely be improved [3].

Sparse Optical Flow was the best performing algorithm across each aggregation method. The best overall result came from sparse optical flow using the median vector to represent the motion of the frame. However, the mean aggregation method provided very similar results, is more efficient to compute, and provides a more representative solution since it indicates the expected value (using the maximum likelihood for a normal distribution) of the motion vectors; assuming few outliers. We choose the mean aggregation method for the remaining experiments for this reason.

Since the dense optical flow algorithm calculates a flow vector for each pixel in the image, and many pixels fall into the non-textured region, the result of the algorithm is to output many flow vectors that are either very small, or the zero vector. Although these vectors are not filtered out of the output, the minimum and even median aggregation methods are choosing these very small vectors that do not accurately represent the actual motion in the frame, explaining the very small amount of variance between the various methods for dense optical flow in Table 1. Assuming there are no errors or poorly chosen features during feature extraction, the sparse optical flow will provide much more meaningful output in this situation.

We also observe that SURF matching performs less accurately in all aggregation methods. A closer look at the SURF matching algorithm output reveals that, while some features were appropriately matched between two consecutive frames, many features in each frame are improperly matched, possibly due to lack of texture in the controlled study design (a sample image of the environment is shown in Fig. 3).

This results in a number of output vectors that are typically much larger than the true values. For example, a feature in the upper left corner of one frame that is matched with a feature in the lower right of the next frame will give a very large flow vector for that point. The existence of these large vectors is clear in the large amount of error for the max aggregation method. Using these large vectors to represent the movement of the frame is inaccurate.

Table 1. Computed SSE Values by Algorithm and Aggregation Method, Vectors, Normalized by Total Frame Count

	Dense Optical Flow	Sparse Optical Flow	SURF Matching
min	53.846	45.212	51.334
median	53.689	**6.421**	1023.548
mean	47.253	**7.080**	178.964
max	4097.331	877.551	489795.918

Sparse optical flow gives a fairly consistent result across most aggregation methods. Due to the selection of a limited number of representative input features, the variance in size and angle of the output flow vectors is very small. For this reason, the choice of aggregation method has lesser impact on the amount of error in the result, since feature selection has eliminated many of the vectors that may become outliers in the final result.

V. Gait Motion Extraction

After extracting the motion vectors from a video, they must be analyzed to detect characteristics of the subject's gait. We turn to signal processing algorithms for this purpose. The predominant factor in determining gait is head motion, since the camera is worn on the head. The application of signal processing measures and algorithms has been shown to be effective in previous work to describe head motion during locomotion [12].

We begin our gait analysis by performing power spectrum analysis on the collected data with the goal of extracting the rate of walking from the video. We first calculate a periodogram from the collected motion vectors. The periodogram aids in identifying the frequency found in the generated motion vectors [16]. Since the gait is constant through our sample video clips, a single constant walking frequency should exist. The peak of the periodogram provides this frequency or the number of cycles in the motion vectors per second, which equates to the number of walking cycles (two steps per cycle, one with each foot) occurring per second.

In order to verify the result, we calculate the periodogram on both the manually collected truth data as well as the motion data generated by finding the mean motion vector from the optical flow algorithm on each frame. A visual inspection of the periodograms shows agreement in each of the 8 collected video clips. The periodograms for each of the four subjects at the two speeds are presented in Fig. 4.

At 2.3 miles per hour, subjects A and B had an identified walking pace of .793 Hz, or one step every 0.63 seconds. At 3.9 mph, the identified rate for both increased to 1.02 Hz, or one step every 0.49 seconds.

Similarly, at 2.3 miles per hour, subject D also had a calculated gait pace of .793 Hz, while subject C had a pace of .963 Hz, or one step every 0.52 seconds. It is expected that subjects C and D may have a quicker walking pace at the same speeds, since they were shorter than subjects A and B and have

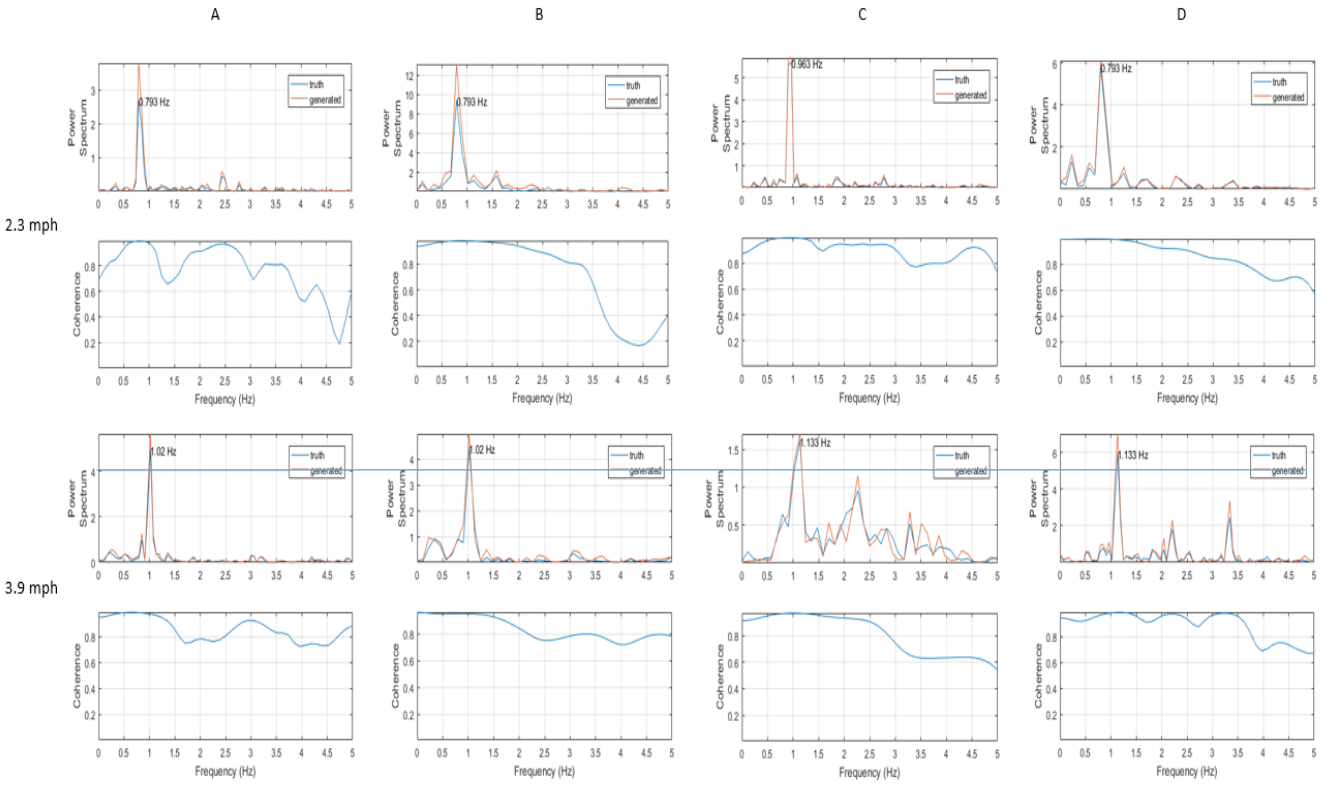

Fig. 4. Periodogram with Generated and Truth Data at 2.3 mph and 3.9 mph for Each Subject, and Corresponding Coherence Function for Generated and Truth Data

smaller stride lengths. This held true at the higher speed, as the step cycle for subjects C and D at 3.9 miles per hour was calculated to be 1.133 Hz, or a step every 0.44 seconds.

To once again formally test the similarity between the manually collected waveform and the automatically generated waveform, as well as compare how well the extracted frequencies in the two match, we performed a magnitude squared coherence measure. The magnitude squared coherence measure indicates similarity between two waveforms as a function of frequency. At frequencies where the waveforms are most similar, we expect the coherence measure to be approaching 1, and be nearer to 0 at frequencies for which the content of the waveforms are dissimilar. That is, if the periodogram calculations have identified similar frequencies in both the truth data and the generated data, we expect to find a high coherence at that frequency, indicative of a good match.

The calculated mean squared coherence results can be found in Fig. 4 (rows 2 and 4 for the eight gait samples). For each of the gait samples, the coherence function illustrates agreement in the truth data and the data generated from the sparse optical flow calculations with a very high accuracy. In each case, the algorithms correctly showed the exact walking pace of each subject that was highlighted in the truth data. This shows a great amount of promise in applying signal processing techniques to data extracted from first person video to analyze gait.

Based on the manual matching of the detected walking cycle frequency data with the true walking cycle frequency from manual annotations, as well as verification of the data against truth data with the coherence function, we have shown that it is possible to reliably extract gait information from a single first person camera sensor in a controlled environment.

VI. Conclusion & Future Work

In this paper, we have proposed a method of detecting and extracting information about gait by collecting data from a single camera sensor embedded in a pair of glasses. We manage to detect gait and extract gait speed in a controlled environment across four different subjects and two different speeds. For our next steps, we plan to expand our gait detection to experiments which include varying speeds in a single video. We also plan to collect data in indoor as well as outdoor settings with a larger number of participants to test the algorithm performance under different illumination conditions, as well as corroborate the results with other wearable fitness trackers that measure similar gait measures. Our preliminary findings show strong promise for using the Pivothead camera for providing gait information in domains of healthcare, rehabilitation, and elder care applications among others.

References

[1] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, Providence, RI, 2012, pp. 2847-2854.

[2] Pivothead. [Online]. Available: www.pivothead.com. Accessed on August 18, 2016.

[3] S. Singh, C. Arora and C. V. Jawahar, "Generic action recognition from egocentric videos," 2015 Fifth National Conference on

Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), Patna, 2015, pp. 1-4.

[4] L. Sun, G. Liu and Y. Liu, "3D Hand Tracking With Head Mounted Gaze-Directed Camera," in IEEE Sensors Journal, vol. 14, no. 5, pp. 1380-1390, May 2014.

[5] B. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAID '81), April 1981, pp. 674-679.

[6] G. Farneback, "Two-frame motion estimation based on polynomial expansion," Proceedings of the 13th Scandinavian conference on Image analysis, 2003, pp. 363-370.

[7] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision 60, 2004, pp. 91-110.

[8] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, SURF: Speeded Up Robust Features, Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, 2008, pp. 346-359.

[9] T. Banerjee, J. M. Keller, M. Popescu, M. Skubic, "Recognizing complex instrumental activities of daily living using scene information and fuzzy logic*," Computer Vision and Image Understanding*, Volume 140, November 2015, Pages 68-82,

[10] D. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz, M. Aud, "Linguistic summarization of video for fall detection using voxel person and fuzzy logic," *Computer Vision and Image Understanding*, Volume 113, Issue 1, January 2009, Pages 80-89

[11] T. Nguyen, N. Jean-Christophe, and F. Francisco, "Recognition of Activities of Daily Living with Egocentric Vision: A Review." Ed. Vittorio M. N. Passaro. *Sensors (Basel, Switzerland)* 2016, 16 (1).

[12] E. Hirasaki, S. T. Moore, T. Raphan, B. Cohen, "Effects of walking velocity on vertical head and body movements during locomotion," Experimental Brain Research, vol. 127, 1999, pp. 117-130.

[13] Y. Watanabe, T. Hatanaka, T. Komuro and M. Ishikawa, "Human gait estimation using a wearable camera," Applications of Computer Vision (WACV), 2011 IEEE Workshop on, Kona, HI, 2011, pp. 276-281.

[14] M. Ermes, J. PÄrkkÄ, J. MÄntyjÄrvi and I. Korhonen, "Detection of Daily Activities and Sports With Wearable Sensors in Controlled and Uncontrolled Conditions," in IEEE Transactions on Information Technology in Biomedicine, vol. 12, no. 1, pp. 20-26, Jan. 2008.

[15] M. Skubic, "Assessing Mobility and Cognitive Problems in Elders," Proceedings, *AAAI Fall 2005 Symposium Workshop on Caring Machines: AI in Eldercare*, Arlington, Virginia, November 4-6, 2005.

[16] F. Auger, and P. Flandrin. "Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method." *IEEE Transactions on Signal Processing*. Vol. 43, May 1995, pp. 1068–1089.

[17] M. Kourogi and T. Kurata, "Personal positioning based on walking locomotion analysis with self-contained sensors and a wearable camera," Mixed and Augmented Reality, 2003. Proceedings. The Second IEEE and ACM International Symposium on, 2003, pp. 103-112.