Temporal Segmentation of Egocentric Videos

Yair Poleg Chetan Arora* Shmuel Peleg
The Hebrew University of Jerusalem
Jerusalem, Israel

Abstract

The use of wearable cameras makes it possible to record life logging egocentric videos. Browsing such long unstructured videos is time consuming and tedious. Segmentation into meaningful chapters is an important first step towards adding structure to egocentric videos, enabling efficient browsing, indexing and summarization of the long videos. Two sources of information for video segmentation are (i) the motion of the camera wearer, and (ii) the objects and activities recorded in the video. In this paper we address the motion cues for video segmentation.

Motion based segmentation is especially difficult in egocentric videos when the camera is constantly moving due to natural head movement of the wearer. We propose a robust temporal segmentation of egocentric videos into a hierarchy of motion classes using a new Cumulative Displacement Curves. *Unlike instantaneous motion vectors, segmentation using integrated motion vectors performs well even in dynamic and crowded scenes. No assumptions are made on the underlying scene structure and the method works in indoor as well as outdoor situations. We demonstrate the effectiveness of our approach using publicly available videos as well as choreographed videos. We also suggest an approach to detect the fixation of wearer's gaze in the walking portion of the egocentric videos.*

1. Introduction

Camera and storage technologies enable to record one's entire day on a camera. Indeed, there is a growing use of wearable cameras which are recording many hours a day. For example, wearable cameras are used in many police districts in order to reduce complaints against policeman (Fig. 1). While recording egocentric video is on the rise, retrieval and indexing of such unstructured videos is still a challenge. Since it is very hard to watch long egocentric videos from start to end, automated tools are needed to enable faster access to the information in such videos.

Figure 1: The use of egocentric cameras is increasing, and are becoming routine in many cases. Examples show such cameras used by policeman, by UN inspectors in Syria, and Google Glass for personal use

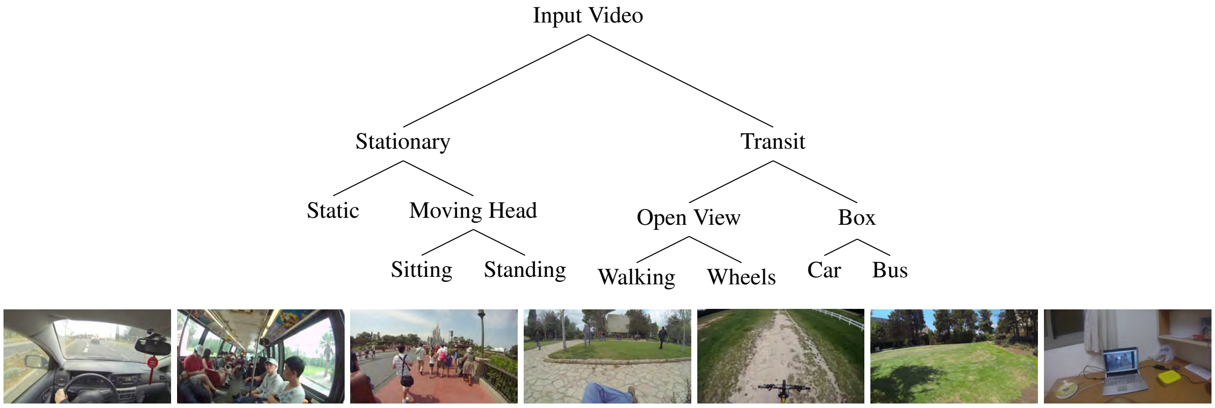
Temporal segmentation adds structure to the video by partitioning the video into chapters. This is a first step for video summarization methods, which should also enable fast browsing and indexing so that a user can quickly discover important activities or objects [8, 9, 11, 15].

There is a growing interest in analyzing and understanding video taken from first person's point of view. Much work addresses understanding objects and activities seen by the camera: hand gestures, object detection or recognition, and activity recognition [5, 14, 16, 18, 19]. These schemes perform well on specialized classes like 'making coffee' or 'applying peanut butter'. Generalizing them to recognize variety of activities, performed routinely by the camera wearer, may involve enormous efforts. Additionally, activities targeted by these approaches usually occur over a short period of time, giving no information about the rest of the video.

Lu and Grauman present in [11] an elaborate analysis and summarization method for egocentric videos. The first step is motion-based temporal segmentation of the video into three classes of *static*, *moving the head*, and *in transit*. Kitani et al. [8] suggest unsupervised learning for video segmentation. The down side of unsupervised learning is that the video segmentation may have no semantic meaning to the viewer. The motion analysis in both these approaches uses instantaneous optical flow which is good for detecting short term activities [1]. This often results in noisy or over

[*] Chetan Arora is now with IIIT Delhi.

[1] By 'short-term' we imply activities which occur over a relatively short period of time, possibly a few seconds, e.g. turning left or sitting



(a) Car (b) Bus (c) Walking (d) Sitting (e) Wheels (f) Standing (g) Static

	Class	Description
Level 1	Stationary	Wearer is at one place. May have natural head motion or even taking a few steps if standing.
	Transit	Wearer is changing location with a forward motion.
Level 2	Static	Head fixated. E.g. sleeping, watching television.
	Moving Head	Stationary with head not fixated.
Level 2	Box	Changing location inside a vehicle with partial view of surroundings.
	Open View	Changing location with full view of surroundings.
Level 3	Sitting	Usual meaning. Natural head motion present.
	Standing	Usual meaning. May be taking a few steps. E.g. waiting for a bus, talking on phone.
Level 3	Walking	Changing location on foot. Slow forward motion. More natural head motion due to stepping.
	Wheels	Faster motion compared to walking. Less head motion. E.g. bicycle.
Level 3	Car	Forward motion with frontal view.
	Bus	Forward motion with sideways view. Riding in the train should also be considered this class.

Figure 2: We suggest temporal segmentation of an egocentric video into hierarchy of 12 classes based upon cues from wearer's head motion. The suggested approach partitions the video into semantically meaningful long term activities (spanning tens of minutes).

segmented partitioning. Markov Random Field (MRF) regularization is used in [11] to smooth the temporal segmentation.

We propose to temporally segment an egocentric video into a hierarchy of activities as shown in Fig. 2. The activities proposed in the hierarchy are long term. Therefore, partitioning with such hierarchy produces macro level segmentation of the video into meaningful chapters. The proposed partitioning can also be used as a pre-processing algorithm similar to the 'pre-process' stage in [11]. The proposed hierarchy can be useful for providing semantically accurate working sets for either activity or object recognition schemes proposed in [5, 14, 16, 18, 19]. For example, algorithms recognizing activities like 'making coffee' should perform better given the information whether the camera wearer is stationary or driving at that time. High-level temporal segmentation of the video can also aid novelty detection and summarization algorithms [3, 9].

The focus of this paper is on analyzing what a wearer does using motion cues due to wearer's activity. However, temporal segmentation of an egocentric video using motion cues poses some key challenges. Though it would have been best to compute egomotion of the camera to find if the wearer is stationary or moving, finding egomotion in egocentric video remains a challenge. Our experiments with Voodoo [2] on egocentric videos indicated that (a) it was hard to use Voodoo on long sequences beyond 1000 frames. (b) the computed egomotion was unstable. The failure can

down. In contrast 'long-term' activities like walking/driving/biking may occur over a period of several minutes/hours

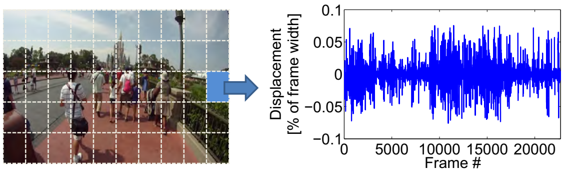


Figure 3: Video frames are divided into a grid of large patches. The instantaneous x-displacement of one patch over 22,000 frames is shown on right. Frames 0-3000 correspond to forward motion which is well hidden in the instantaneous displacements.

be attributed to dominant 3D rotation induced by the head motion leading to degenerate epipolar geometry. Feature tracking is harder and unreliable in such videos due to significant depth variations and dynamic objects in the scene, making egomotion computation further difficult. Experiments with VisualSFM [1], PTAM [4] and our own implementation of egomotion estimation led to similar conclusions.

We propose a new tool for long-term temporal segmentation, the *Cumulative Displacement Curves*; replacing the commonly used feature tracking and instantaneous motion analysis. Using integrated motion instead of instantaneous motion allows us to focus on long term activities. Additionally, it makes our method stable and robust against natural head motion and moving objects in the field of view. The tool is generic and can be used to recover hierarchy of activity classes as shown in Fig. 2.

We also suggest in this paper an approach to recognize fixation of the camera wearer's gaze using egocentric video. In [7], gaze is inferred using an eye-tracking camera to improve recognition rate. Eye-tracking along with camera ego-motion are used in [13] to recognize indoor activities. In [10], gaze is estimated from the camera wearer's head motion and hand location. While eye tracking is not possible from egocentric video, we observe that gaze fixation is often coupled with 'head fixation'. We therefore detect such head fixation from egocentric video as evidence for gaze fixation. Our method detects fixation of gaze in 'walking' sequences from the outward looking camera.

Inertial measurements and a wearable camera are used by [17] to temporally segment recorded motions into activities. Our method on the other hand uses only the recorded video, which by definition is always available in egocentric video and works both in indoor and outdoor environments. It is obvious that the combination with external sources, such as GPS and inertial sensors, may give better results.

Organization of the paper is as follows. Sec. 2 describes *Cumulative Displacement Curves* and a technique to generate such curves. Sec. 3 describes the classification of wearer's motion using the cumulative displacement curves. In Sec. 4 a method to find gaze fixation in walking segments of the egocentric video is presented. Experiments are given in Sec. 5 using publicly available videos [6] as well as videos shot by us. The source code and the dataset are available at the project page: *http://www.vision.huji.ac.il/egoseg/*

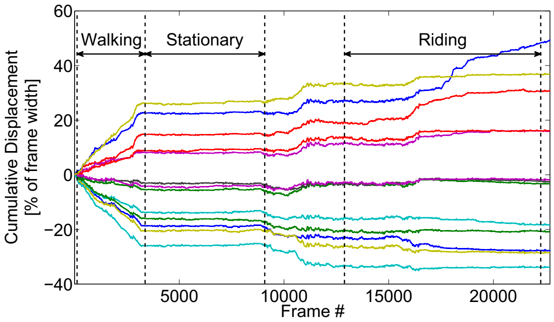


Figure 4: Cumulative displacement curves for some patches shown in Fig. 3. Time periods with expanding curves, as in frames 1-3000, correspond to 'walking'. Periods where the cumulative curves are horizontal correspond to 'standing'. Periods with some horizontal and some expanding curves correspond to driving.

2. Cumulative Displacement Curves

Most methods that compute camera ego-motion start with the detection and tracking of feature points from frame to frame. Displacement of these features is used to find the camera motion. Large head rotations make the tracking of feature points very difficult in an egocentric video. We found that computing displacements at fixed image patches is more efficient and robust. We divide the image into a grid of $W \times H$ cells (10×5 in our implementation), and compute the displacements:

$$d_t(i,j) = (d_t^x(i,j), d_t^y(i,j)) \tag{1}$$

of each cell $(i,j) \in (W \times H)$ at time t, where $d_t^x(i,j)$ and $d_t^y(i,j)$ denotes the x and y displacements of the cell (i,j) (see Fig. 3). We refer to each cell as *motion detector*.

In pure forward motion the displacement at any image location is outwards from focus of expansion. In an egocentric video with the camera wearer moving forward, the instantaneous displacement has large variations caused by the head rotation (see Fig. 3). However, over a long period of time the average of the displacement caused by head rotation is practically zero. Correspondingly, integrating the instantaneous displacements results in canceling out of the zero mean variations due to head rotation, leaving only the consistent displacement caused by forward motion. We denote the integration of instantaneous displacements as:

$$D_t(i,j) = \sum_{k=1}^{t} d_k(i,j) \tag{2}$$

and refer to it as *cumulative displacement* or *CD* in short.

Fig. 4 shows the plot of the x component of the *CD* against time, which we call the *CD* curve. Measuring trends in the *CD* curves allows us to focus on long term activities while ignoring small perturbations due to head motion. Further, occasional failures in optical flow calculation have negligible effects on these trends.

When walking, which is always forward, the *CD* curve increases for patches on the right of the FOE, and decreases for patches on the left of the FOE. When stationary, the *CD* does not change, and the *CD* curve is horizontal for all patches. These cases are shown in Fig. 4.

An interesting set of *CD* curves is obtained when riding inside a vehicle (car or a bus). When a person is driving a car, some parts of the frame show the outside scene visible through the window. The *CD* curves for these parts are increasing and decreasing as expected from forward motion. Other parts of the frames show the inside of the car. *CD* curves corresponding to inside locations will indicate a stationary wearer with horizontal *CD* curves. So when some *CD* curves are horizontal, while others have upward or downward trend, this is an indication for driving. This situation can happen also when a wearer is sitting in a bus looking sideways (the y *CD* curve may be horizontal in this case but the x *CD* curve would still have large slope). We repeat that we are looking at the long-term trends, therefore texting on a phone or talking to somebody for a few seconds and not looking outside have little effect on such pattern.

3. Classification of Wearer's Motion

As described in the previous section, the trends in the *CD* curves can be used to predict the wearer's activity. The slope of the *CD* curve represents the instantaneous motion of the patch. We are interested in 'long-term' activities spanning minutes and by implication the long term trends in *CD* curves. We therefore smooth the *CD* curves by convolving them with a Gaussian kernel of a large σ, and use the slopes of the smoothed curve as a measure of long-term trends. Mathematically:

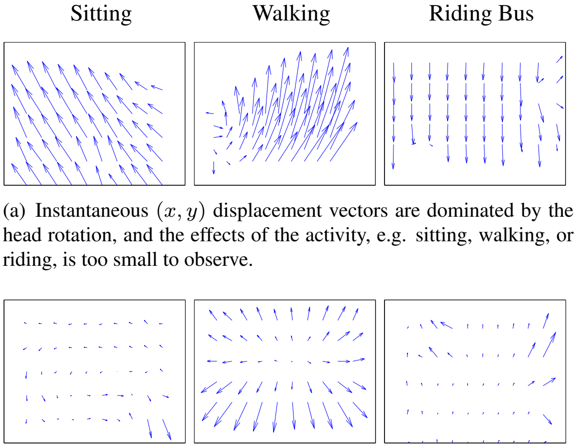
$$M_t(i,j) = \frac{\partial}{\partial t}\left(D_t(i,j) * N(0,\sigma)\right), \qquad (3)$$

where $M_t(i,j) = (m_t^x(i,j), m_t^y(i,j))$ is the *motion vector* for cell (i,j) at time t. We observe that M_t can be equivalently found by smoothing the original displacement vectors $d_t(i,j)$:

$$M_t(i,j) = d_t(i,j) * N(0,\sigma). \qquad (4)$$

Although gaussian smoothing worked well for our case, bilateral filtering or other approaches could have been used as well in order to find long-term trends in the *CD* curves.

Fig. 5(b) shows samples of motion vectors corresponding to stationary, walking and riding frames. Note that the plot of the motion vectors is very much as expected from "clean" motion: radially outwards when 'walking', partially radially outwards when riding a 'bus'/'car' and very small when 'sitting'. These plots are much more characteristic of the global ego-motion, compared to the instantaneous displacements in Fig. 5(a) which are dominated by the natural head rotation. The process of building the *CD* curve and measuring only long term trends removes the noise of head rotation, enabling robust inference of the wearer's long-term activity.



(a) Instantaneous (x, y) displacement vectors are dominated by the head rotation, and the effects of the activity, e.g. sitting, walking, or riding, is too small to observe.

(b) Motion vectors obtained from the cumulative displacement curves as given in Eq. 3. Effects of head rotations are removed, and the direction of vectors are now noiseless. For 'walking' the vectors are large and have radial direction. In the 'sitting' case, the magnitude mostly zero. Riding ('car'/'bus') has a mixed pattern.

Figure 5: Comparing instantaneous displacement vectors and motion vectors obtained from cumulative displacement curves

The σ of the Gaussian used for smoothing the *CD* curve in Eq. 3 controls the shortest event that can be detected. In our experiments we use the Gaussian smoothing corresponding to a time resolution of roughly 17 seconds. This implies that if a person is walking for 10 minutes and stops in between for more than 17 seconds, the proposed scheme should output walking-standing-walking segmentation. Stopping for less than 17 seconds will result in a single walking segmentation. An implication of the proposed approximation is the reduced accuracy on the activity boundaries. The Gaussian smoothing filters the high frequency components representing activity boundaries in a *CD* curve. Smoothing such edges results in mixed slopes at the boundaries, thereby reducing the classification accuracy of the algorithm on the activity boundaries.

After obtaining the motion vectors, we train SVM classifiers for each binary classification in the proposed class hierarchy (see Fig. 2). These classifiers use various features derived from motion vectors. We describe these features below.

3.1. Radial Projection Response

It is expected that in forward motion the motion vectors (as shown in Fig. 5(b)) will point radially outwards from focus of expansion. To test if M_t follows such a pattern, we perform the following steps. We threshold $M_t(i,j)$ by its magnitude and normalize it to a unit vector:

$$\widehat{M_t}(i,j) = \begin{cases} \frac{M_t(i,j)}{\|M_t(i,j)\|}, & \text{if } \|M_t(i,j)\| \geq \tau \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

We set τ to be a small number (in our implementation $\tau = 0.002$) in order to ignore insignificant motion vectors. Next, we project all motion vectors, $\widehat{M_t}(i,j)$, on a template containing unit vectors pointing out radially from the focus of expansion:

$$R_t(i,j) = \frac{C_{i,j} - FOE}{\|C_{i,j} - FOE\|} \cdot \widehat{M_t}(i,j) \quad (6)$$

where $C_{i,j}$ is the (x,y) coordinates of the center of cell (i,j) and and FOE is the location of focus of expansion. Note that $R_t(i,j)$ is the cosine of the angle between the motion vector $\widehat{M_t}(i,j)$ and a vector pointing outwards from the focus of expansion, through the center of cell (i,j). Finally, we obtain the score S_t by counting the number of motion vectors $\widehat{M_t}(i,j)$ whose directions are within angle ϕ of the corresponding vector in the radial projection template:

$$\widehat{R}_t(i,j) = \begin{cases} 1, & R_t(i,j) \geq \cos(\phi) \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

$$S_t = \sum_{(i,j)} \widehat{R}_t(i,j). \quad (8)$$

We call S_t the *Radial Projection Response*.

For a stationary wearer (static/sitting/standing) the motion vectors are very small, and the radial projection response is low. The radial projection response is high when the wearer is in transit ('box' or 'open view'). Using the radial projection response to distinguish between moving and stationary is much more robust than using the magnitude of the motion vectors alone.

3.1.1 Estimating the FOE Location

Since we do not know the location of focus of expansion (FOE) a-priori, and since the FOE location may even change during the sequence, we need to compute the FOE. For each time instance independently, we search several locations, and select as the FOE the location which maximizes the *radial projection response* at that time instance.

For efficiency reasons we restricted the candidate FOE locations to the center 50% of the frame's area. This works well for walking, as in most cases the viewing direction of the camera is forward. For the stationary case, where no significant FOE exists, choosing any candidate near the center works OK.

3.2. Motion Clusters

The *radial projection response* (Eq. 8) can be used to classify the video segment as stationary or transit. The *radial projection response* is also low during riding, since there may be only a few motion detectors outside the vehicle indicating a moving pattern. What can distinguish riding from stationary is the fact that when stationary most motion detectors have small motion vectors, whereas in riding the motion vectors corresponding to outside regions have large magnitude. Therefore, we sort the motion vectors by magnitude and compute the average magnitude of top 6% and bottom 6% of motion vectors. The two averages and their difference are used as features in the SVM classifier.

3.3. Statistical Information

We also use the following per frame statistical information to help in the classification: Number of motion detectors with successful instantaneous displacement computation, average magnitude of the motion vectors (x and y separately) as obtained from Eq. 3 and average and standard deviation of the instantaneous displacements as given by Eq. 1. This information is available per frame and is appended to the feature vectors used by SVM classifiers.

4. Detecting Period of Gaze Fixation

In a forward camera motion all *CD* curves should have upward or downward trend (depending upon location of the motion detector) with a fixed slope corresponding to wearer's speed. The curve for the walking portion can be approximated well by a smoothed version of the curve as described in previous section. Because of the left and right motion of wearer's head, the original curve correspondingly moves above and below the smooth approximation respectively. This natural motion of the head temporally stops when a wearer's gaze is fixated on something. The event is visible as all *CD* curves remaining temporarily below or above their smooth approximations. We detect this anomaly to identify gaze fixation and thereby important segments in the walking portion of video. Figure 6 explains the process.

We consider the area between the original curve and the smoothed curve as positive when the original curve is above the smoothed curve and negative when below. The absolute value of the area is proportional to both the magnitude of head movement and the time period for which it remains left or right with respect to the forward viewing direction. We find the *cumulative difference curve* by integrating the signed area. As a wearer keeps moving his head equally to

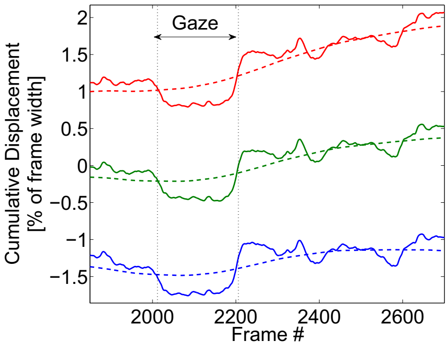


Figure 6: Original (solid) and smoothed (dashed) cumulative displacement curves of 3 selected motion detectors, for the selected frame range in a test sequence. Because of the left and right motion of wearer's head, the original curve correspondingly moves above and below the smooth approximation respectively. The periodic alternation of the head left and right temporally stops when the gaze is fixated on something. In a cumulative displacement curve the effect is seen as original curve remaining temporarily below or above the smooth approximation. The same can be detected and used to identify gaze fixation.

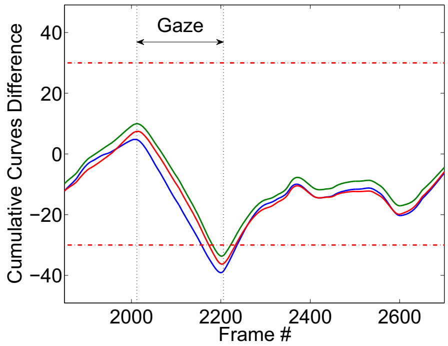


Figure 7: Cumulative difference curve corresponding to cumulative displacement curves in Fig. 6. The two dashed lines represent the threshold beyond which the motion detector fires. The gaze fixation has been detected at around frame 2200.

the left and right, the cumulative difference curve has globally horizontal trend near zero, with top and bottom peaks corresponding to head crossing the forward viewing direction after going left and right respectively. When the gaze is fixated on something, we see significantly higher peaks in cumulative difference curves (see Fig. 7). We identify such extraordinarily high peaks in each cumulative difference curve separately (corresponding to each motion detector). The corresponding motion detector is then flagged as supporting a gaze fixation hypothesis. If more than a certain number of motion detectors supports the fixation hypothesis, the frame is classified as part of gaze fixated segment. It may be noted that there can be spurious peaks in a individual cumulative difference curve due to moving objects (especially close objects) in the scene. While gaze fixation is visible at all motion detectors, peaks from moving objects will not be supported by all motion detectors, and can be removed by voting.

	Accuracy			# Samples	
Classifier	**Avg.**	**Class** 1	**Class** 2	**Class** 1	**Class** 1
Box vs. Open	93%	94%	92%	176K	853K
Car vs. Bus	74%	73%	75%	121K	58K
Sitting vs. Standing	70%	72%	67%	499K	453K
Static vs. Moving	96%	98%	94%	25K	999K
Stationary vs. Transit	90%	86%	93%	992K	1053K
Walking vs. Wheels	93%	96%	91%	778K	126K

Table 1: Classification results for each of the binary classification tasks. Class 1 in the classifier refers to prior class in the name of classifier. E.g., in a 'Box vs. Open' classifier, class 1 refers to 'Box' and second to 'Open'. We show class-wise accuracy from which binary confusion matrices can be derived.

5. Experiments

We tested our algorithm on a dataset of videos shot at Disney World [6] and videos downloaded from YouTube. Along with these we have videos captured by us using GoPro cameras. In all there are more than 65 hours of video. All videos are shot from a head-mounted camera by 13 subjects in various locations, performing all sorts of activities, under various lighting conditions and times of the day. The test videos captured by us are with resolution of 1280×720 at 30 frames per second (FPS). Other videos (Disney and YouTube) are with various resolutions at FPS of either 15 or 25. We have manually annotated the videos taken from YouTube and the ones captured by us. Disney videos come with annotations which we have updated for our experiments. The videos captured by us along with our annotations for the complete dataset is available at the project page: *http://www.vision.huji.ac.il/egoseg/*. We have also released the source code for our implementation at the project page.

Frame to frame instantaneous displacements were computed using LK [12] on the patches of a 10×5 grid. motion vectors were normalized for frame size of 1×1. In all experiments, we fixed $\tau = 0.002$, $\phi = \pi/2$. We flagged a motion detector as supporting gaze if the cumulative difference is more than a standard deviation away from its mean. We flag the frame as gaze if more than 80% motion detectors support the gaze hypothesis. All experiments ran on a desktop PC.

5.1. Motion Classification

To evaluate the accuracy of our motion classification we labeled each video frame as one of the seven leaf-nodes in the graph shown in Fig. 2. We labeled frames to be excluded from the experiment as 'DontCare'. This is typi-

	Walking	Car	Standing	Bus	Wheels	Sitting	Static
Walking	**83%**	0%	6%	6%	4%	1%	0%
Car	1%	**74%**	3%	15%	0%	3%	4%
Standing	14%	2%	**47%**	4%	0%	31%	2%
Bus	3%	19%	27%	**43%**	0%	7%	1%
Wheels	9%	0%	0%	6%	**86%**	0%	0%
Sitting	3%	1%	22%	1%	0%	**62%**	10%
Static	0%	1%	1%	0%	0%	1%	**97%**

Table 2: Confusion matrix for the cascaded classifier tree. Rows are ground truth. Diagonal elements represents class accuracy, off diagonal elements give pairwise confusion.

Class Label	Accuracy	# Samples
Static-Moving	91%	1083115
Sitting-Standing	82%	1036217
Box-Open	87%	1197623
Car-Bus	76%	228108
Walking-Wheels	82%	969515

Table 3: Inner nodes of the hierarchy have important semantic meaning. Therefore we give the accuracy of the cascaded classifier considering each inner node as a class label by itself.

cally for completely dark frames where tracking failed completely. For the Disney videos of [6], we manually corrected their annotations at a few places where it was not accurate enough for our needs. For example when in a train, while the original annotation referred to as train ride, we changed the annotation to 'sitting' when the train stops for a long period of time.

We train a separate SVM classifier for each binary classification task represented by an inner node in the classification hierarchy shown in Fig. 2. As features, we concatenate the *radial projection response*, number of blocks with successful LK evaluation, all motion vectors, top and bottom motion vector cluster centroid and their distance, instantaneous displacement average and standard deviation (for static-moving classifier only). We have a dataset of 140 sequences (including Disney, YouTube and the ones captured by us), from which we choose sequences at random for training. We keep choosing a sequence at random until we get 12500 training samples/frames for each class. The remaining sequences are used for testing. Table 1 gives the accuracy obtained for each of the binary classification task independently. Table 2 gives the confusion matrix of the cascaded classifier, where a classifier down the hierarchy process only the frames classified as positive by its parent. For example, Static-Moving classifier operates only on the frames marked 'stationary' by Stationary-Transit classifier.

While Table 2 is important from the perspective of knowing the classification performance at the leaf nodes of the hierarchy, we consider inner nodes of the hierarchy equally important. For example its useful to know that we can classify accurately when a wearer is transiting in 'open' or in 'box', even if we can't accurately tell whether by a 'car' or a 'bus'. Table 3 gives the accuracy of the cascaded classifier at the inner nodes. The results compare favorably with the ones obtained in the related area [7], though there is no direct comparison between the two approaches. The work of Kitani et al. [8], though targeting temporal partitioning, is not directly comparable to ours, owing to its unsupervised nature and focus on short term atomic actions. It may be noted that we haven't employed any smoothing on the classifier results and every frame is classified separately using the features described above. We expect that using a regularization framework like MRF on the classification results, as done by Lu and Grauman [11], may further improve the results.

We observed that the classifier accuracy is markedly better for the sequences in which activities happen on a time range of several minutes. One of the explanation for this is the choice of blurring for segmenting the *CD* curves resulting in undesired smoothing on the activity boundary. When the activity itself is long, this has relatively little effect. On the other hand if the activity occurs over a span of few seconds, the blurring results in unwanted mixing of features from temporally adjacent activities leading to reduced accuracy of the classifier. This is particularly visible in the 'Sitting vs. Standing' classifier. Many of the sequences that we have for riding scenario have intermittent stoppages which are labeled as 'sitting'. The feature vector in this short term sitting gets affected by the temporally adjoining activities which is mostly riding in our sequences. This is one of the reasons for the relatively weaker performance by the 'Sitting vs. Standing' classifier.

Another limitation of the proposed classification scheme is what we refer to as 'mixed activities'. For example, the proposed scheme is able to distinguish between 'standing' and 'walking' quite accurately. However, there may be some semantically meaningful activities like 'waiting' which involves a mix of stationary and walking and does not fit atomically into the proposed framework. Fig. 8 shows some other failure cases.

5.2. Detecting Gaze Fixation

The publicly available dataset [6] did not have any segments for gaze fixation. We therefore tested our strategy for detecting gaze fixation on 'walking' sequences captured by us. The ground truth annotations were done manually. We consider a fixation of the wearer's head for more than 5 seconds as a valid gaze fixation. Table 4 lists results achieved on various test sequences. The proposed scheme is able to

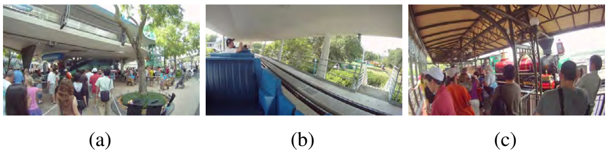


Figure 8: Activity classification failures. (a) Waiting in line, a mix of standing and walking. Our algorithm fails to handle mixed activities. (b) Riding an open train. The pattern of the approximated motion field is very close to the 'open' class pattern. (c) Standing while the train is coming into the station. While some of the frame is static, other parts are moving fast and therefore the frame is classified as 'box'.

Seq.	**Frames**	**# Fixation Detected**	**# True Positives**	**Accuracy**
C1-C2	32017	47	39	82.97%
Y1-Y8	121208	219	163	74.43%
Total	153225	266	202	75.93%

Table 4: Gaze Fixation Results on sequences taken by two subjects. Total of 10 sequences were tested.

detect more than 75% fixation instances correctly.

Some of the failure cases that we have observed arise from ambiguity in gaze fixation as observed from head motion. For example, a left and right turn in quick succession leads to similar 'bumps' in the cumulative difference curve as observed during the gaze fixation. The same bump is observed if a person turns in place. There could have been couple of heuristics to improve upon such cases. For example, it is expected that during a gaze fixation at an interesting object, the person would try to see it again and again. Such repeated 'short' gaze would cause consecutive bumps in the cumulative difference curves and can be used to distinguish gaze from large head motion because of other reasons. We leave such improvisations for future work.

6. Conclusion

Temporal segmentation of an egocentric video into 12 hierarchical classes (7 disjoint classes) is presented. The proposed classification hierarchy partitions the video such that semantically meaningful inference can be made for every frame of the video. Unlike prior approaches, we focus on long time activities preventing over-segmentation of the video. Wide variations in the scene and movement of wearer makes any inference in the egocentric video a challenging task. Use of cumulative displacement curves allows us to model long-term activity patterns which have been used for the temporal segmentation and detecting gaze fixation. Classification using feature vectors derived from the proposed cumulative displacement curves is simple, efficient and robust against local tracking failures. The important shift in the focus from what user sees to what user does has led to a highly accurate classifier. We expect similar head movement patterns to be distinct in other activities like working on computer, writing or washing dishes. The same therefore can be generalized to detect such targeted short-term activities as well. We leave such improvements for the future work.

Acknowledgement: This research was supported by Google and by the Israel Ministry of Science.

References

[1] VisualSFM : A Visual Structure from Motion System, Changchang Wu, http://ccwu.me/vsfm/. 3
[2] Voodoo Camera Tracker, Digilab, http://www.digilab.uni-hannover.de/. 2
[3] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty detection from an ego-centric perspective. In *CVPR*, 2011. 2
[4] R. O. Castle, G. Klein, and D. W. Murray. Video-rate localization in multiple maps for wearable augmented reality. In *IEEE ISWC*, 2008. 3
[5] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011. 1, 2
[6] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012. 3, 6, 7
[7] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012. 3, 7
[8] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011. 1, 7
[9] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 1, 2
[10] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013. 3
[11] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 1, 2, 7
[12] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 2, 1981. 6
[13] K. Ogaki, K. M. Kitani, Y. Sugano, and Y. Sato. Coupling eye-motion and ego-motion features for first-person activity recognition. In *CVPR Workshops*, 2012. 3
[14] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 1, 2
[15] Y. Pritch, A. Rav-Acha, and S. Peleg. Nonchronological video synopsis and indexing. *PAMI*, 30(11), 2008. 1
[16] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013. 1, 2
[17] E. Spriggs, F. D. L. Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *CVPRW*, 2009. 3
[18] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *PAMI*, 1998. 1, 2
[19] S. Sundaram and W. Mayol-Cuevas. High level activity recognition using low resolution wearable vision. In *CVPRW*, 2009. 1, 2