Understanding the Nature of First-Person Videos: Characterization and Classification using Low-Level Features

Cheston Tan, Hanlin Goh, Vijay Chandrasekhar, Liyuan Li, Joo-Hwee Lim
Institute for Infocomm Research, Singapore
{cheston-tan, hlgoh, vijay, lyli, joohwee}@i2r.a-star.edu.sg

Abstract

First-person view (FPV) video data is set to proliferate rapidly, due to many consumer wearable-camera devices coming onto the market. Research into FPV (or "ego-centric") vision is also becoming more common in the computer vision community. However, it is still unclear what the fundamental characteristics of such data are. How is it really different from third-person view (TPV) data? Can all FPV data be treated the same? In this first attempt to approach these questions in a quantitative and empirical manner, we analyzed a meta-collection of 21 FPV and TPV datasets totaling more than 165 hours of video. We performed the first quantitative characterization of FPV videos over multiple datasets, encompassing virtually all available FPV datasets. Validating this characterization, linear classifiers trained on low-level features to perform FPV-versus-TPV classification achieved good baseline performance. Accuracy peaked at 81% for 2-minute clips, but 67% accuracy was achieved even with 1-second clips. Our low-level features are fast to compute and do not require annotation. Overall, our work uncovered insights regarding the basic nature and characteristics of FPV data.

1. Introduction

With many consumer devices having wearable cameras coming to market, videos recorded from a first-person view (FPV) are set to become increasingly common. However, it is still unclear what exactly characterizes or defines FPV (also termed *ego-centric*) videos and differentiates them from videos taken from a third-person view (TPV).

Take for instance a school play: the same scene might turn out quite differently, depending on whether it was recorded by a parent casually wearing a Google Glass device, or a parent using a top-of-the-line camcorder with image stabilization and trying to make it "look good". The same semantic content could look either FPV-like or TPV-like. Conversely, the same device can produce videos with either FPV-like or TPV-like characteristics, such as professional camera crews intentionally adjusting their camerawork for a sense of intimacy and immediacy during "reality" TV shows, or during regular TV shows with FPV-like portions (*e.g.* simulating the killer's point-of-view during a crime drama). In other words, neither semantic content nor recording device alone clearly define whether a video is perceived as FPV or TPV.

Why is characterizing and understanding FPV data important? In the field of computer vision, FPV or egocentric vision is an increasingly popular topic, and this will only continue as FPV video content becomes more common. However, fundamental questions have been left unanswered – they have not even been asked. What makes FPV data different, such that new algorithms may be required? Is all FPV data the same, such that an algorithm validated on one FPV dataset will work well on other FPV datasets? In this work, we take the **first steps in trying to address these questions in a quantitative, empirical manner**.

While there are various intuitive notions of what FPV-like characteristics are, there has not been a rigorous, quantitative characterization of FPV videos across multiple datasets. These intuitions include low-level factors such as the long and unstructured nature [15], rapid changes [13] and variation [23] in illumination, significant camera motion [13] and motion blur [23]. Factors related to higher-level semantic information include close views of objects [23], complex hand-object interactions [13], and hand occlusion [23] and proximity [12].

These intuitions may not necessarily be accurate. Chest-mounted cameras are not affected by head movement, so there is no jerky motion due to looking around. Cameras mounted on top of the head (or cap or helmet) may not capture body parts belonging to the wearer, another supposed FPV characteristic. Only by actually **implementing these intuitions as features and then testing on real data**, can the question of FPV characteristics be properly answered.

On a more practical note, can videos be automatically classified as FPV or TPV with good accuracy? This is a novel problem that may become increasingly important. Certain algorithms are designed specifically for FPV videos (*e.g.* [15]). Also, for video-hosting sites like YouTube, if FPV videos can be detected, they could undergo post-processing or enhancement for better viewing. Conversely, one might want to intentionally post-process TPV videos to give them a more authentic, first-person feel.

Our paper has 3 main contributions: (1) We provide the first quantitative characterization of FPV videos across multiple datasets – in particular, across nearly all the FPV datasets that are currently publicly available. (2) We show that a set of simple, low-level features that are easy to compute and do not require annotation for training, are sufficient for a baseline/benchmark accuracy of 81% in distinguishing FPV and TPV clips. (3) We uncover specific features and insights about FPV data – what makes it different, and whether it is homogeneous.

In the rest of this paper, we review related work, then describe the datasets and features used. We show that these low-level features can discriminate FPV and TPV video clips well. Then, we analyze specifically which features discriminate FPV/TPV best, and why. Finally, we examine the differences among FPV datasets. Overall, this initial work provides insight into the nature of FPV data.

2. Background

First-person view (FPV) visual processing is gaining ground in the computer vision community. Recent years have seen more and more FPV papers and datasets (see Table 1). In a short number of years, FPV data has been studied for a wide range of topics, including discovering, detecting and recognizing people [8, 12], hands [13], objects [7, 22, 23] and activities [9, 20, 26, 27]. Other topics include social interactions [5], video summarization [12, 15] and novelty detection [2].

Despite various qualitative intuitions about FPV characteristics (see Section 1), little work has been done to investigate the nature of this data quantitatively. One exception is the "early" work by Ren & Philipose [23]. Analyzing their FPV dataset of 42 day-to-day objects being manipulated in the course of daily activities, they studied the challenges and constraints of their dataset, such as motion blur, hand occlusion, location prior and temporal consistency. Some interesting findings include their estimate of drops in SIFT-based recognition accuracy of 20% due to background clutter and 13% due to hand occlusion.

We go beyond this important initial work by examining many more FPV (and TPV) datasets, in an attempt to perform a characterization not based solely on one dataset. This scaling-up presents a challenge for some of the analyses in [23], which required time-consuming manual annotation. Rather than annotating the more than 165 hours of video contained in the datasets, we instead focused solely on low-level features requiring no annotation. As the rest of this paper will show, these features can go surprisingly far.

3. Methods: datasets, features and classifier

As part of our goal of characterizing and classifying FPV videos, we assembled a large collection of 21 FPV and TPV datasets (see Table 1). To the best of our knowledge, the 13 FPV datasets included comprise the vast majority of FPV datasets available for public download (with the exception of one recent dataset [1]). Because there are many existing TPV datasets, our choice of TPV datasets is necessarily an idiosyncratic sampling. Nonetheless, we attempted to maximize diversity in terms of content (*e.g.* short action clips, surveillance videos, daily activities) and source (*e.g.* Hollywood, the internet, created by computer vision researchers). Our choice of datasets is in no way a statement regarding quality or popularity of these datasets (or excluded ones).

	Dataset	Vids	Hr	Cam	Comments	
01	CMU	171	17	Head	Kitchen	[4]
02	Disney	113	50	Head	Disneyland	[5]
03	GTEA	28	.5	Head	Indoors	[7]
04	Gaze	17	1	Head	Indoors	[6]
05	Gaze+	30	5	Head	Kitchen	[6]
06	IEOR	10	2	Lapel	Indoors	[23]
07	JPL	57	.5	Head	Worn by toy	[26]
08	EgoADL	20	10	Chest	Daily activs.	[20]
09	UEC	2	.5	Head	Outdoors	[9]
10	UTE	4	17	Ear	Life-logs	[12]
11	UTokyo	5	2	Head	Office	[18]
12	W31	31	2	Collar	Walking	[2]
13	YouTube	6	.3	Var.	Outdoors	[9]
14	HMDB	6766	8	–	Movie clips	[10]
15	Hwood	475	2	–	Movie clips	[11]
16	Hwood2	2517	9	–	Movie clips	[16]
17	UCF50	6677	17	–	YouTube	[21]
18	URADL	155	1	–	Daily activs.	[17]
19	UTI	20	.5	–	Interaction	[25]
20	VIRAT	329	9	–	Surveillance	[19]
21	Shows	11	13	–	Movies, TV	–

Table 1. Datasets 01 to 13 are FPV, while datasets 14 to 21 are TPV. *Vids*: number of videos in dataset. *Hr*: total duration of dataset in hours. *Cam*: camera placement. *Var.*: various placements.

Because the FPV datasets on average contained videos that were many times longer than the typical short clips in TPV datasets, we attempted to somewhat balance this by creating our own mini-dataset consisting of 11 full-length movies and TV episodes (duration ranging from 42 minutes to almost 3 hours). This is the *Shows* dataset in Table 1.

In total, the datasets contained more than 165 hours of video (approximately 107 hours FPV and 59 hours TPV) covering a diverse set of activities, objects and locations. The TPV datasets included a few without any camera motion (*URADL*, *UTI* and *VIRAT*). The FPV datasets were collected using both head-mounted cameras and other more statically-mounted (*e.g.* chest-mounted) cameras. The FPV videos were collected during activities ranging from walking to work (*W31*), food preparation (*CMU*, *GTEA*, *Gaze* and *GazePlus*), object manipulation (*IEOR*), first-person interaction (*JPL*), office activities (*UTokyo*), daily activities (*EgoADL*), outdoors sports (*UEC* and *YouTube*), an outing to Disneyland (*Disney*) and unconstrained activities (*UTE*). The durations ranged from up to 5 hours (*UTE*) to around one minute or less (*GTEA*). Some FPV datasets were collected purely indoors and with little whole-body movement, while others included walking from indoors to outdoors (and vice-versa). Furthermore, several FPV datasets used wide-angle lenses. **In short, the FPV datasets contained considerable variation, making any attempt to characterize FPV data as a whole very challenging.**

3.1. Dataset standardization

Because the 21 datasets also came in vastly different resolutions, aspect ratios and frame rates, we standardized them. The standardized resolution was 320x240 (aspect ratio of 1.33). Video frames were first resized to a height of 240 pixels, and then cropped to a width of 320 pixels. Videos that were originally shorter than 240 pixels or had an aspect ratio less than 1.33 were discarded. There were 766 discarded videos from 3 TPV datasets (*HMDB*, *Hollywood* and *Hollywood2*), less than 5% of their original number.

Next, to serve the dual purpose of standardizing the frame rate and reducing the amount of data, frames were extracted from videos at 1 fps (frame per second). This was the highest common factor among the various frame rates (*i.e*. 15, 24, 25, 30 and 60 fps), and also because the *W31* dataset was only available for download as still frames extracted at 1 fps. Nonetheless, even at this low frame rate, there were more than 500,000 frames in total.

3.2. Features

We operationalized some commmon intuitions about FPV characteristics (see Section 1) as simple, low-level features (see Table 3). As a start, we restricted ourselves to three classes of features: blurriness, illumination and optical flow. FPV videos might contain more head and body movement, leading to certain characteristics relating to optical flow (*e.g*. left-right motion due to looking around), while such movement may also lead to more motion blur. Another intuition is that FPV videos may contain more indoor-to-outdoor transitions (and vice-versa), and unlike movies, such drastic illumination changes are not adjusted for.

We chose relatively simple and fast algorithms for computing blurriness, illumination and optical flow features, not necessarily the latest or most accurate algorithms. For blurriness, we followed the work of Lu & Grauman [15] and used the algorithm of [3]. For optical flow, we used the SIFT flow algorithm of Liu *et al*. [14, 15]. Instead of passing adjacent frames (spaced 1s apart) to the algorithm, we extracted additional video frames 200 milliseconds after every frame (which are at 1 fps). For videos at 24 fps, we used the frame closest to 200 millisecond; the error is less than 5%. For the *W31* dataset, which was only available at 1 fps, we did not compute optical flow. For both blur and optical flow algorithms, we used the default parameters. For illumination, we simply used the mean or median pixel intensity in a video frame as a simple proxy for illumination.

From these 3 sets of basic low-level features, we computed a total of 50 features. Broadly speaking, each feature is computed by performing a series of operations on the blur, illumination or optical flow features. Examples of operations include computing the mean over the entire frame, computing the standard deviation over all frames in a video or clip, and z-scoring (normalizing to mean 0 and standard deviation 1) across all frames in a video. Other operations include computing the first- and second-order temporal derivatives, approximated by simply finding the difference between adjacent frames (and repeating the process for second-order derivative).

Apart from these generic operations, blur and optical flow had other operations. The blur algorithm returned two numbers for each frame (vertical and horizontal blur). We also took the max over these two numbers, in order to summarize the blurriness of a frame into a single number.

For optical flow, the SIFT flow algorithm returns a dense optical flow estimate, *i.e*. an x and y flow estimate for every pixel in the frame. We converted these into magnitude and angle, and then quantized the angles into 8 bins, centered at 0°, 45° and so on. These angles were then summarized over frames using three different methods: 1) the number of times each angle bin contained the most motion energy in a frame (*ANG-nrg*), 2) the number of times each angle bin was the most common bin (*i.e*. the mode) in a frame (*ANG-mode*) and 3) a total count of pixels belonging to each angle bin (*ANG-count*). For each method, values were ultimately normalized so that the sum over all 8 bins was 1.

Overall, each feature essentially summarized one entire video or clip into a single number. Of the 50 features we computed, 24 were related to optical flow angle (3 methods x 8 angle bins), 8 were related to optical flow magnitude, 6 were related to illumination, and 12 were related to image blur. Due to space constraints, we cannot describe all 50 features in detail here, but the naming convention in Table 3 is relatively self-explanatory, given the descriptions in this section. As an example, one of the more informative illumination features involved first taken the median pixel intensity over all pixels in a frame, and then computing the standard deviation of that median value over all frames in a video. This feature is denoted as *ILLU-med-stdev*, and corresponds to the intuitive notion of how much the global illumination varies over the duration of a video. Another example is the feature *BLUR-max-stdev*; the max over horizontal and vertical blur values is first computed for each frame, and then the standard deviation over all frames in a video is computed. This corresponds to the intuition notion of how much the blurriness varies over a video (perhaps a rough gauge of how much start-stop head motion there is).

3.3. Classifier

We used the Regularized Least Squares (RLS) classifier with a linear kernel, which has been shown empirically to have similar performance to Support Vector Machines (SVMs) on several datasets [24]. We used 1e-10 as the regularization parameter throughout. The data was z-scored (whitened), such that each feature had a mean of 0 and standard deviation of 1, before being passed to the classifier. This was done separately for training and test data.

Due to the highly imbalanced number of examples from the positive (FPV) and negative (TPV) classes, the training examples from the class with more examples was randomly sub-sampled, such that ultimately the classifier was trained on an equal number of positive and negative examples. This "balancing" was validated by chance-level performance when training labels were randomly shuffled as a control; this was performed whenever a classifier was trained. Without balancing, shuffling of training labels resulting in performance significantly greater than chance.

4. How well can FPV and TPV clips be discriminated?

In this section, the aim is to verify that the 50 features can in fact separate FPV and TPV clips with a reasonable level of accuracy. We examine classifier performance when all videos are first chopped into clips of fixed length, before classifier training and testing proceeds. There are two reasons for this. The first is that videos can be arbitrarily long (*e.g.* some videos in the *UTE* dataset are up to 5 hours long), so it would be highly desirable if only a short clip were needed to determine if the whole video is FPV or TPV.

The second reason is more pragmatic. Some of the TPV datasets are in fact comprised of short clips extracted from longer videos, leading to thousands of short "videos"; on the other hand, few (or none) of the videos in the FPV datasets were similarly treated. Thus, chopping all videos up into fixed-length clips makes for a fairer comparison.

Note, however, that the distinction between FPV and TPV may become more ill-defined as clip length shortens. For example, given only 1-second clips, it is unclear how well humans are at differentiating FPV and TPV. Some clips could be rather ambiguous, *e.g.* an outdoor mountain scene with only a little motion, which could either be from a TPV documentary or an FPV video from someone's travels.

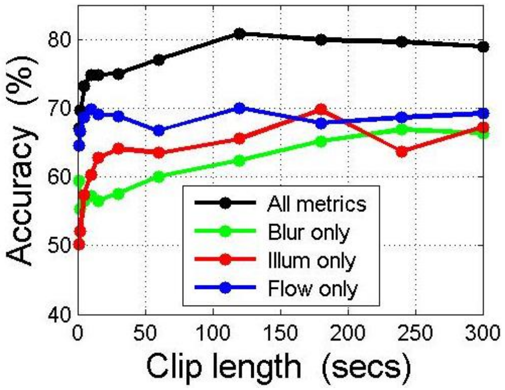

Figure 1. FPV/TPV classifier accuracy as a function of clip length.

We examined classifier performance for clip lengths ranging from 1 second to 5 minutes (see Fig. 1). A clip length of 120 seconds gave the best accuracy of **80.9%**, while performance declined very slightly for longer clip lengths, reaching **79.0%** at 5 minutes. As expected, performance drops as clip length is reduced from 120 seconds to 1 second. Interestingly, however, even for clips as short as 1 second, performance is still significantly above chance, at **67.1%**. As a sanity check, regardless of clip length, performance was at chance when labels were shuffled during training. In the rest of this paper, we use a clip length of 120 seconds unless otherwise stated.

These results indicate that **even with simple features and a linear classifier, a good baseline accuracy of over 80% can be obtained**. The regime of very short clip lengths is of particular interest, since the amount of computation required can be two orders of magnitude less (*e.g.* 300 seconds vs. 1 second). **Baseline performance for 1-second clips was well above chance, at over 67%**.

4.1. Feature classes

Among the three classes of features (blur, illumination and optical flow), how do they perform when used in isolation as input features for the classifier? As Fig. 1 shows, the optical flow features (blue) were best overall, while the blur features (green) were worst overall. At all clip lengths, blur features performed worse than optical flow features. However, the maximum accuracy for each of the classes was close: **70.0%** for optical flow, **69.8%** for illumination, and **66.9%** for blur. These occurred at different clip lengths, so combining the feature classes at their own optimum clip lengths could in theory boost overall accuracy. For very short clips, optical flow alone performed almost as well as all three features combined, suggesting that the overall combined performance is driven primarily by optical flow.

4.2. Generalization to unseen datasets

So far, while training and test sets are distinct, datasets under test have also been part of the training set. In practice, because of the diverse nature of videos "in the wild", we cannot assume that a new video has similar characteristics as some pre-trained dataset. This also links to the issue of dataset bias; given that the FPV datasets in this paper can be quite diverse, what happens when entire datasets are completely unseen during training? We train a FPV/TPV classifier as before, training on half the clips in each dataset and testing on the other half. The only difference here is that we remove all clips of a particular dataset from the training set, and put them in the test set. This is done in turn for each dataset. For each dataset, we compare the performance when it was unseen by the classifier during training, to the mean performance during times when it was seen.

Averaged across datasets, accuracy when unseen during training of the classifier was **67.1%**, compared to **77.8%** (a **10.7%** drop in accuracy; see Table 2). This is a relatively small drop in performance. However, some datasets suffered large drops in performance. In particular, accuracy on the FPV datasets *CMU* and *UTE* dropped by **21.0%** and **22.4%** respectively, while the TPV dataset *VIRAT* suffered a drastic **64.3%** drop. These results suggest that while certain datasets (such as *VIRAT*, which is the only surveillance dataset used in this paper) may suffer catastrophically, overall there is decent generalization to unseen datasets, supporting the broad generality of our features.

	Dataset	Seen	Unseen	Drop
01	CMU	47.8	26.8	21.0
02	Disney	72.3	61.9	10.4
03	GTEA	–	–	–
04	Gaze	69.4	68.1	1.3
05	Gaze+	83.6	79.7	3.9
06	IEOR	94.5	94.2	0.3
07	JPL	–	–	–
08	EgoADL	46.1	30.3	15.8
09	UEC	98.1	99.2	-1.1
10	UTE	59.7	37.3	22.4
11	UTokyo	41.1	33.1	8.0
12	W31	97.0	96.6	0.4
13	YouTube	82.1	82.6	-0.5
14	HMDB	–	–	–
15	Hwood	–	–	–
16	Hwood2	99.8	99.8	0.0
17	UCF50	–	–	–
18	URADL	–	–	–
19	UTI	–	–	–
20	VIRAT	98.8	34.5	64.3
21	Shows	99.2	95.1	4.1
	Mean	77.8	67.1	10.7

Table 2. Accuracies when datasets are part of the training set (*"seen"*) or not (*"unseen"*), and the drop in accuracy. Note that even for the *"seen"* datasets, the reported accuracies are for clips not part of the training set. For some datasets, there were insufficient clips of 120 seconds in length; accuracies are marked –.

5. What makes FPV data different from TPV data?

In the previous section, we found that collectively, our 50 features show reasonably good performance on a FPV/TPV classification task. This shows that together, the features are able to capture the characteristics that distinguish FPV data. However, are there specific features that characterize FPV data better than others features?

In this section, we examine each feature separately, in order to pinpoint which features characterize FPV data well. Similar to the previous section, we trained linear classifiers to perform FPV/TPV classification (see Section 3.3 for details). The only difference is that we used each feature separately, *i.e.* there were 50 classifiers instead of 1.

Table 3 shows the accuracies using each of the 50 features separately. The overall best feature is *ANG-nrg-1*, which achieves **77.2%** accuracy. This is surprisingly close to the accuracy of **80.9%** when using all 50 features, and it is by far the best feature (next best accuracy is **71.8%**). In other words, one very characteristic feature of FPV data is the motion energy in the rightward direction.

It is important to note that many of the 50 features produce accuracies close to chance, indicating that **the problem itself can be very challenging.** Of the 50 features, 16 of them (almost 1 in 3) have less than 55% accuracy. This strongly indicates that beyond the very rough intuition that blur, illumination and motion are important FPV characteristics, it is **far from trivial to implement these intuitions as good features – the details of each feature are crucial**.

5.1. Blur features

Among the 12 features related to blur, the *BLUR-hor-stdev* feature (standard deviation of horizontal blur among the frames in a clip) gives the highest accuracy of **66.1%**, while **66.0%** is attained by *BLUR-max-z-d2-ab-mean* (mean absolute 2nd-order rate-of-change of omni-directional blur). Together with *BLUR-max-z-d1-ab-mean* achieving **65.5%** accuracy, the results support the intuition that **changes in blur due to acceleration (or change in acceleration) are characteristic of FPV data.**

While this finding makes sense, this is not an *a priori* obvious result. Many of the TPV clips include human actions – with natural bursts of acceleration and deceleration – which could also cause changes in blur. In fact, 4 of the 8 TPV datasets are human action recognition datasets, with clips specially cropped from longer videos or movies to focus on actions (*i.e.* static scenes removed). As such, only a quantitative, empirical study such as this would shed light on how well change in blur really characterizes FPV data.

Indeed, it is important to note that blur by itself is not a strong distinguishing characteristic of FPV data. Among the features relating to the average amount of blur, the highest accuracy is only **54.5%** (*BLUR-max-med*, the median

Feature	Accu	Feature	Accu
BLUR-hor-med	54.0%	ANG-nrg-1	**77.2%**
BLUR-hor-mean	53.2%	ANG-nrg-2	60.4%
BLUR-hor-stdev	**66.1%**	ANG-nrg-3	56.1%
BLUR-ver-med	51.2%	ANG-nrg-4	57.9%
BLUR-ver-mean	49.2%	ANG-nrg-5	65.4%
BLUR-ver-stdev	61.9%	ANG-nrg-6	60.1%
BLUR-max-med	54.5%	ANG-nrg-7	62.4%
BLUR-max-mean	49.6%	ANG-nrg-8	61.5%
BLUR-max-stdev	60.3%		
BLUR-max-z-d1-ab-mean	65.5%	ANG-mode-1	65.9%
BLUR-max-z-d2-ab-mean	66.0%	ANG-mode-2	61.6%
BLUR-ratio-med	55.6%	ANG-mode-3	52.9%
ILLU-med-stdev	51.0%	ANG-mode-4	53.8%
ILLU-mean-stdev	59.0%	ANG-mode-5	**71.1%**
ILLU-mean-d1-ab-mean	52.4%	ANG-mode-6	57.8%
ILLU-mean-d2-ab-mean	53.9%	ANG-mode-7	57.4%
ILLU-mean-z-d1-ab-mean	**69.5%**	ANG-mode-8	55.2%
ILLU-mean-z-d2-ab-mean	66.1%		
MAG-d1-ab-med-med	62.7%	ANG-count-1	**71.8%**
MAG-d1-ab-mean-mean	56.3%	ANG-count-2	65.8%
MAG-med-med	63.9%	ANG-count-3	59.8%
MAG-mean-mean	60.3%	ANG-count-4	46.1%
MAG-mean-z-d1-ab-mean	66.8%	ANG-count-5	69.5%
MAG-mean-z-d2-ab-mean	**68.2%**	ANG-count-6	59.1%
MAG-mean-d1-ab-mean	51.5%	ANG-count-7	53.4%
MAG-mean-d2-ab-mean	51.1%	ANG-count-8	54.8%

Table 3. Accuracy values for all 50 features. The best features for each class or sub-class are in **bold**. The features beginning with MAG and ANG are related to optical flow magnitude and angle respectively. Clip length is 120 seconds.

amount of omni-directional blur in a clip). This is despite the fact that for many types of TPV data (*e.g.* internet clips, TV shows, movies), either the cameras are controlled to avoid jerky motion, or the videos are post-processed to remove jerky motion. This might suggest in theory that TPV data contain significantly less blur than FPV data, but in reality the numbers do not support this.

5.2. Illumination features

Among the 6 illumination-related features, the best-performance 2 features are *ILLU-mean-z-d1-ab-mean* and *ILLU-mean-z-d2-ab-mean* (the mean absolute 1st- and 2nd-order rate-of-change in illumination, where the illumination of a frame is approximated as the mean pixel intensity). The accuracies are **69.5%** and **66.1%** respectively. Note that because each dataset (or even video) may have different illumination characteristics (*e.g.* indoors or outdoors), it is important to normalize for overall mean and standard deviation. Without this step, the above two accuracies drop to less than **54%** (*ILLU-mean-d1-ab-mean* and *ILLU-mean-d2-ab-mean*).

The finding that rate-of-change in illumination is characteristic of FPV data is actually surprising. Most of the FPV datasets are either predominantly indoors or predominantly outdoors. The primary exception is the *UTE* life-logging dataset, which is very unconstrained and diverse. Even then, transitions from indoors to outdoors (and vice-versa) are relatively rare.

As such, it is unlikely that the classifier performance is driven by drastic changes in global illumination. More likely, it is due to micro-changes in mean pixel intensity due to camera or object motion. Again, because the TPV datasets also contain actions, it is not *a priori* obvious that a single feature related to micro illumination changes can produce almost 70% classification accuracy.

Importantly, it is not so much the overall variation in illumination (*ILLU-mean-stdev*, **59.0%**), but rather the **short timescale (frame to frame) micro-changes in illumination that are more characteristic of FPV data.**

5.3. Optical-flow magnitude features

Similar to the features related to illumination, for the features related to optical-flow magnitude, the best performing two features compute the mean absolute 1st- and 2nd-order rate-of-change (*MAG-mean-z-d1-ab-mean* and *MAG-mean-z-d2-ab-mean*; **66.8%** and **68.2%** respectively. Again, normalization is important; without normalization, accuracies drop to less than **52%** (*MAG-mean-d1-ab-mean* and *MAG-mean-d2-ab-mean*).

The overall median and mean amount of optical flow (*MAG-med-med* and *MAG-mean-mean*) do not perform as well (**63.9%** and **60.3%** respectively). In other words, **FPV clips generally do have more optical flow than TPV clips, but this is not as strong a characteristic as the rate-of-change of optical flow** (with the mean amount of optical flow normalized away).

5.4. Optical-flow angle features

For the optical-flow features for specific directions, there is a consistent trend across all 3 sets of 8 directions. Within all sets, the leftward and rightward directions (1 and 5) produce the highest accuracies. Moreover, these produce some of the highest accuracies among all 50 features. In other words, **horizontal motion statistics are among the strongest characteristics of FPV data**, even beyond overall amount of motion or change in motion. **This aspect of motion has not been previously mentioned as a key characteristic of FPV data**, illustrating the value of performing a comprehensive, quantitative analysis.

Somewhat counter-intuitively, it is not that FPV data has more horizontal motion than TPV data. One might expect more horizontal motion due to left-right head movement, *e.g.* when looking around. However, as Fig. 2 shows, it is TPV data that has more horizontal motion. Apart from the outlier FPV datasets *JPL* and *UTokyo* (also see Section 6), all the TPV datasets have very large proportions of optical

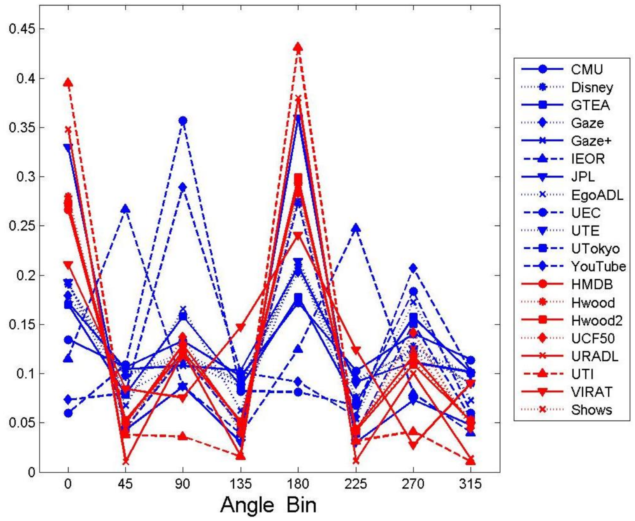

Figure 2. Distribution of direction of optical flow for FPV (blue) and TPV (red) datasets. Right is 0° and left is 180°.

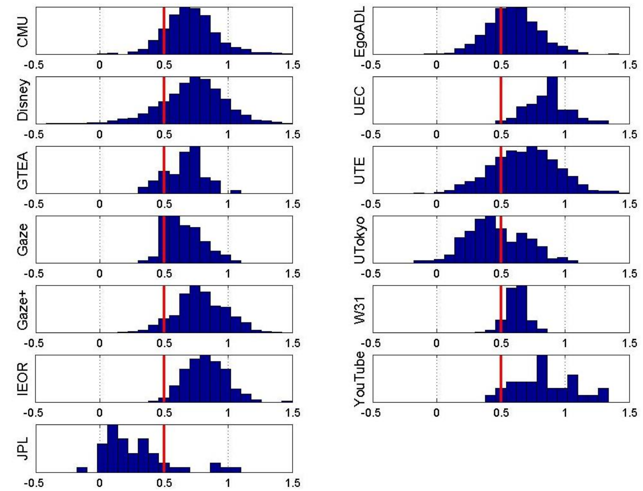

Figure 3. Histograms of raw classifier values (x-axis) for FPV datasets. The y-axis represents the number of clips. Since only the distribution of values is of interest, the y-axis of each sub-plot are independently scaled for presentation purposes. Clips with values above 0.5 (right of the red line) are classified as FPV.

flow in the 0° (right) and 180° (left) directions, compared to other directions. With the benefit of hindsight, this can be explained as most TPV motion being due to horizontal camera panning or motion along the ground plane.

6. Is all FPV data the same?

Thus far, we have treated all the FPV datasets as a single category, by classifying FPV clips against TPV clips. In this section, we compare the various FPV datasets to see whether they are relatively homogeneous, or if in fact they can be quite different from one another. This is an important question. For instance, if FPV datasets are actually extremely heterogeneous, then one cannot claim to have developed an algorithm for FPV data in general, unless the algorithm is validated on a variety of FPV datasets.

Here, in order to compare FPV datasets, we again train a linear classifier to perform FPV/TPV discrimination, but then examine the raw classifier output (*i.e*. the weighted sum of the 50 features). We plot the output for each dataset as a histogram (see Fig. 3), with the output value on the x-axis, and the number of clips having such a value on the y-axis. A clip with classifier output value of more than 0.5 is classified as FPV (see Table 4). We use a clip length of 10 seconds rather than 120 seconds (74.8% rather than 80.9% accuracy in Fig. 1), so that all FPV datasets have a reasonable number of clips, making the histograms more representative. To make full use of the data, all clips are put into the training set; there is no left-out test set. (Here, we are not interested in generalization accuracy; we are interested in how homogeneous FPV datasets are.)

It is important to compare FPV datasets in a quantitative, data-driven manner, and not just compare the qualitative differences of the datasets (*e.g*. whether the data is collected indoors or outdoors). Without a data-driven comparison, it is unclear whether any supposed differences actually matter in terms of what really characterizes FPV data.

As the results from Fig. 3 and Table 4 show, FPV datasets are not homogeneous. There are 2 clear outliers in terms of the distribution of raw classifier values (and the resulting accuracy). The *JPL* and *UTokyo* datasets have accuracies of **14.6%** and **46.7%** respectively. Clearly, these datasets are somehow different from other FPV datasets, and their clips are often mis-classified.

Why might this be the case? One unique difference about the *JPL* dataset [26] is that the camera was worn by a teddy bear, not an actual person. There is camera motion induced by moving the teddy bear, but this was achieved by pushing it around on an office chair. As such, there is no jerky, human-like head or body motion, even though this dataset is clearly egocentric in the sense that the camera records humans interacting with the camera-wearer. The fact that this results in the *JPL* dataset achieving only **14.6%** accuracy further confirms the findings of Section 5.

The causes are less clear-cut for the *UTokyo* dataset [18], which was collected using head-worn cameras while people performed typical activities in an office setting. One possible reason is that there were periods of little movement while people looked at the computer screen. Again, this is consistent with jerky motion being a key FPV characteristic. However, this drives home the fact that TPV-like segments can exist even with unambiguously FPV recording setups, and suggests that **subjective human perception of FPV/TPV may be a better definition of ground-truth.**

	Dataset	Accu		Dataset	Accu
01	CMU	87.5	12	W31	97.0
02	Disney	83.5	13	YouTube	93.5
03	GTEA	85.2			
04	Gaze	88.7	14	HMDB	80.4
05	Gaze+	92.5	15	Hwood	78.3
06	IEOR	98.7	16	Hwood2	76.5
07	JPL	14.6	17	UCF50	68.4
08	EgoADL	68.8	18	URADL	90.4
09	UEC	100.0	19	UTI	98.5
10	UTE	77.9	20	VIRAT	90.0
11	UTokyo	46.7	21	Shows	80.9

Table 4. Accuracies corresponding to histograms in Figure 3. Clips with values larger than 0.5 were classified as FPV. Accuracies for TPV datasets (14 to 21) are also reported, for completeness.

7. Future work and discussion

This is a first attempt at quantitative characterization and classification of FPV videos over multiple datasets. There are many future possibilities, *e.g*. dividing the image into sub-regions, which may allow for discovery of motion patterns corresponding to walking, sitting, *etc*.

Overall, we have shown that a set of simple, low-level features related to blur, illumination and optical flow are able to characterize the differences between FPV and TPV videos, quantified by a classifier accuracy of more than 80%. This baseline using low-level features and a linear classifier is likely to be improved upon using more elaborate or high-level features and more sophisticated classifiers.

One main insight is that rapid changes (rather than variation *per se*) are characteristic of FPV data, clarifying prior intuitions. However, examination of two atypical but unambiguously FPV datasets raised a deeper issue. Even in the absence of rapid changes, there are higher-order characteristics of FPV data that are due to egocentricity itself, rather than due to the camera being worn by people.

References

[1] Workshop on Wearable Computer Vision Systems 2013. 2

[2] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty detection from an ego-centric perspective. In *CVPR 2011*. 2

[3] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas. The blur effect: perception and estimation with a new no-reference perceptual blur metric. In *Electronic Imaging 2007*, pages 64920I–64920I–11. 3

[4] F. De la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, and J. Macey. Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database. 2

[5] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *CVPR 2012*. 2

[6] A. Fathi, Y. Li, and J. M. Rehg. Learning to recogize daily actions using gaze. In *ECCV 2012*. 2

[7] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*. 2

[8] T. Kanade and M. Hebert. First-Person Vision. *Proceedings of the IEEE*, 100(8):2442–2453, Aug. 2012. 2

[9] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR 2011*. 2

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011. 2

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR 2008*. 2

[12] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR 2012*. 1, 2

[13] C. Li and K. M. Kitani. Pixel-level hand detection in egocentric videos. In *CVPR 2013*. 1, 2

[14] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. SIFT flow: dense correspondence across difference scenes. In *ECCV 2008*. 3

[15] Z. Lu and K. Grauman. Story-Driven Summarization for Egocentric Video. In *CVPR 2013*. 1, 2, 3

[16] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR 2009*. 2

[17] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV 2009*. 2

[18] K. Ogaki, K. M. Kitani, Y. Sugano, and Y. Sato. Coupling eye-motion and ego-motion features for first-person activity recognition. In *CVPR 2012 Workshop on Egocentric Vision*. 2, 8

[19] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*. 2

[20] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR 2012*. 2

[21] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2012. 2

[22] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR 2010*. 2

[23] X. Ren and M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *CVPR 2009*. 1, 2

[24] R. Rifkin, G. Yeo, and T. Poggio. Regularized Least Squares Classification. In Suykens, Horvath, Basu, Micchelli, and Vandewalle, editors, *Advances in Learning Theory: Methods, Model and Applications*, pages 131–154. 2003. 4

[25] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, 2009. 2

[26] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR 2013*. 2, 7

[27] E. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *CVPR 2009*. 2