University of Glasgow | School of Computing Science

Learning to Recognise Actions in Egocentric Video

Peter Le Bek

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

Level 4 Project — March 28, 2014

Abstract

Motivated by the potential for head-mounted devices to automatically launch applications based on the users current action, we show that it is possible to recognise everyday actions (e.g. running, writing, idling) in egocentric video footage using motion-based features alone. Using an approach based on the popular HOF motion descriptor and other features specially suited to egocentric video, we achieve a classification accuracy of 82.67%. We introduce a new dataset to address the lack of an egocentric dataset for everyday actions.

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: ______________________ Signature: ________________________

Contents

Chapter 1

Introduction

1.1 Goal

The goal of this work is to recognise everyday actions (e.g. running, writing, idling) in first-person video footage. We are particularly motivated by the potential for head-mounted devices to automatically launch applications based on action context. For example, when the subject starts running the device could automatically display metrics like speed and distance travelled. This is a stepping stone toward more interactive context-awareness: applications that can understand the subjects actions at a fine-grained level and give feedback that prompts new actions.

1.2 Background

Recognizing human activities in video is a well-studied problem in computer vision [19, 11]. Research in this area may use any combination of fixed cameras in the scene, sensors attached to limbs and joints, head-mounted cameras, and head-mounted gaze trackers. In a strictly *egocentric* vision setting we allow only for a head-mounted camera, directed such that it approximates the vision of the subject. This subproblem has received relatively little attention from the research community, and yet we believe it deserves special attention for the following reasons: (1) Many state-of-the-art vision techniques are inspired by the retina, which itself evolved in an egocentric setting (the eye); (2) Wearable technology is entering the consumer space, and there's much interest in utilizing the front-facing camera for computer vision applications; (3) Egocentric vision has unique properties that simplify certain aspects of action recognition:

- **Consistent Perspective** One of the difficulties of recognizing actions with a fixed-camera in the scene is objects and humans look and move very differently depending on the perspective. Egocentric vision does not suffer from this problem because objects and subject motion always appear from the same perspective. Approaches to action recognition in the fixed-camera setting address this problem by designing complex features that attempt to be invariant to perspective, or by training on a very large dataset that includes examples at many different perspectives.
- **Close and Unobstructed Perspective** For actions that involve objects, egocentric vision represents the most detailed, unobstructed view of the discriminative region in the scene. Since humans manipulate objects in front of themselves, egocentric vision generally gives a close, frontal look at any objects of interest. This is a huge win over cameras fixed in the scene where body parts generally obstruct objects in use.

- **Personalized** In the egocentric setting the subject of the action is fixed, whereas in the fixed-camera setting action recognition is usually expected to be robust to various subjects. The natural variation in human height, build and gait, can produce very idiosyncratic motion patterns. By fixing the subject we're free to learn these idiosyncrasies rather than attempt to generalize.

1.3 Motivating Application

The application that inspired this work is that of automatically launching applications on head-mounted devices. This is desirable because it removes the need for explicit interaction, thus saving time and freeing the hands or voice to perform the action itself. By removing the tedium of application launching the subject can conveniently use applications more often.

In this work our goal is to build an action recognition system that could support this kind of behaviour. Thus our system should at minimum be able to report the start and end of an action for a given set of actions. This would make it trivial for the OS to automatically launch and exit the appropriate application.

1.4 Other Applications

Action recognition has other interesting practical applications besides the support of context-aware applications.

1.4.1 Healthcare

Patient-monitoring is becoming more important as the population ages. One of the key measures used by medics to determine the functional status of a patient, particularly in the disabled or elderly, is ability to perform activities of daily living (ADL). In the absence of 24-hour care this is difficult to measure. Egocentric action recognition offers a practical and economically viable solution to this problem. A head-mounted camera coupled with an action recognition system of the kind developed in this work could produce a comprehensive ADL log from the moment a patient wakes up to the moment they go to bed. This log could be monitored by medics from afar, removing the need for an expensive house visit.

1.4.2 Mobile Worker Tracking & Training

Advances in mobile devices have enabled the workforce to become more mobile. Action recognition would allow mobile workers to automatically generate a work log, which could be tracked in realtime from corporate headquarters. This kind of application is particularly apt to labour that requires both hands such as welding or lab work, since the proposed system can be operated hands-free.

Head-mounted action recognition systems could also be used to train mobile workers on new tasks, whereby the worker would progress through a sequence of *action checkpoints*: recognized actions that prompt the next step in the task. When a task is too rare or complex to be learned, the action recognition system could take complete control, directing and recognizing each action toward some end goal.

1.4.3 Improving accessibility for blind and visually-impaired people

Head-mounted action recognition could be used as part of a system that aids the blind or visually-impaired with everyday tasks that are difficult to perform without sight. For example, we can recognize when the users opens the fridge and launch a food recognition system. Action recognition is an important component in a system like this because it's a strong indicator of intent. A single head-mounted camera is significantly cheaper than installing fixed cameras in every room in a house.

1.5 Challenges

1.5.1 Lack of Training Data

The most popular public datasets in the area of action recognition are Hollywood2 and KTH. Neither of these datasets are egocentric, and the egocentric action recognition datasets that do exist are not general purpose enough for our needs. This presents two problems: (1) We lack training data to build our model; (2) We cannot compare egocentric action recognition results with the large body of non-egocentric results.

We address the first of these problems by building our own dataset. This is time consuming both from a planning and implementation perspective. It further increases the complexity of our work because we must argue for the difficulty of recognition tasks within our dataset, whereas the difficulty of the Hollywood2 dataset, for example, is well established.

The second problem of comparison with previous results cannot be solved so easily. Ideally the author would like to see a new action recognition dataset in which egocentric cameras, fixed cameras, gaze trackers, and body sensors, are all employed to provide a multi-faceted view into each action example. With this new dataset previous results could be reproduced, and would now be comparable across all popular settings. Unfortunately building such a dataset is outside the scope of this work.

1.5.2 Indirect Perspective

The egocentric setting is interesting for action recognition because the subject is mostly invisible. Hands and forearms are often visible, and sometimes they contain clues, but in many cases we have to learn about the subject without ever seeing the subject. The brain is well adapted to this - when sitting in the passenger seat of a car we are not required to look at the dashboard or see a video of the car from the outside to know that it is moving.

1.6 Approach

Our approach uses motion-based features to represent egocentric actions. Learning is performed using an SVM.

1.6.1 Feature Extraction

Motion-based features are extracted from pairs of consecutive frames in the input video (Figure 1.6.1). We estimate motion in the scene using Farneback optical flow [6]. From this we extract Histograms of Optical Flow

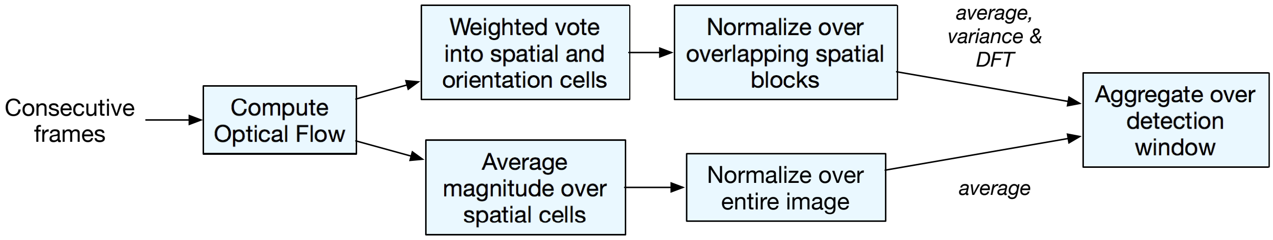

Consecutive frames
Compute Optical Flow
Weighted vote into spatial and orientation cells
Normalize over overlapping spatial blocks
Average magnitude over spatial cells
Normalize over entire image
average, variance & DFT
average
Aggregate over detection window

Figure 1.1: Feature extraction process for our action recognition approach.

(HOF), our main motion descriptor. We also extract the magnitude image and features that characterize the frequency component of motion. Our features are aggregated over a fixed-length time window resulting in a 5105 dimensional feature vector.

1.6.2 Learning

A multi-class SVM, trained on 75 minutes of training video, is used to predict the action class of unlabelled video.

1.7 Acheivements

We evaluate our approach on 15 minutes of continuous test video containing 4 actions on which the model has been trained and several more actions on which it hasn't. Our approach achieves an overall classification accuracy of 82.67%.

1.8 Thesis Statement

The author believes that by leveraging the unique properties of egocentric video, an egocentric action recognition system appropriate for context-aware application launching on head-mounted devices can be built using only optical flow and derived features.

Chapter 2

Background

This chapter contains a review of the action recognition literature and an examination of existing action recognition datasets. We conclude with a refined vision for this work.

2.1 Overview

Our goal is to build an egocentric action recognition system that performs well for both indoor and outdoor activities. What follows is a review of the action recognition literature. We select papers that cover state-of-the-art approaches for a diverse set of actions types (e.g. indoor, outdoor, high-intensity, low-intensity, multi-step). Since the majority of the research in this area takes place in a fixed-camera setting, our review necessarily includes much of this – in addition to as much as we could find on egocentric action recognition.

2.2 Previous Work

2.2.1 Action Recognition

Action recognition has a rich literature which spans many divergent subproblems. Much of the research is focused on recognizing activities from the perspective of a fixed camera in the scene [5, 19, 7, 17, 12], and indeed the most popular datasets in this area (KTH, Hollywood2) are of this nature.

2.2.2 Optical Flow Interpretation

Efros et al. [5] use optical flow to recognise the activities of sports players at a distance. They develop a new motion descriptor based on the highly successful static image descriptor Histogram of Oriented Gradients (HOG), and call it Histogram of Optical Flow (HOF). The optical flow field is split into a dense array of cells, and within each cell optical flow angles are voted into oriented bins, weighted by the local gradient magnitude. The authors claim this global approach is resistant to a moving background or camera, and capable of discriminating between activities at very low resolution. Quantitative results are not provided, but Dalal et al. go on to use HOG/HOF with great success to recognise humans in video [3].

Global HOF descriptors are not robust to variation in recording conditions, since motion patterns depend on the location of the subject in the frame. Additional moving objects cluttering the frame can further hamper the effectiveness of global features. To overcome these problems, Laptev et al. [19] develop an approach based on local space-time features. On the KTH dataset, performance of an SVM classifier trained on local features is shown to be significantly better than that trained on global HOF features. Later research performed by Wang et al. [20] shows this result doesn't hold for more realistic datasets like UCF sports and Hollywood2, where dense (global) sampling performs better than local features. This shortcoming is attributed to the low quality of local spatio-temporal features detected in realistic video.

2.2.3 Action Modelling

More recent work has moved the focus from optical flow interpretation to understanding how human activities should be modelled. The hope is to build models with deeper understanding of human activities, for example the ability to recognise multi-stage or simultaneous activities.

Determining when an action starts and ends in a video has been the topic of numerous papers [18]. It's an important problem because training a classifier on anything but the activities (versus what comes in between them) will reduce its accuracy. Satkin et al. [18] define an action as the most discriminative portion of a video, which gives rise to a natural start and end. Classification accuracy on ideally cropped videos is found to give a significant improvement over uncropped video.

Moving beyond the temporal cropping problem, state-of-the-art methods instead use multi-scalar sliding windows [4]. This approach has practical significance since most real-world video input to recognition tasks is not temporally cropped. The latest research [7] abandons the sliding window in favour *actoms*, short portions of a clip centred on highly discriminative regions and sized according to proximity to nearby actoms. Actoms are composed to form complete activities. This approach has two nice properties: (1) resistance to variations in speed and duration - enabled by the ability of actoms to partially overlap and the adaptive nature of the actom itself; (2) the ability to model activities that contain gaps.

Many human activities are distinguishable by the objects they involve, thus an entirely different approach to motion-based models is to model the objects instead [16]. A major benefit of using objects rather than optical flow is that objects with discriminative significance don't necessarily move. For example, when washing hands at a sink the tap handles don't move when the hands are under the water. Object-based models deal with this quite naturally since the taps are detected in the scene and assimilated into the model whether they move or not, methods based on optical flow must overcome this problem by increasing the complexity of their temporal model. The object-based model presents two problems: (1) objects in the scene aren't necessarily involved with the action, and just confuse the classifier; (2) real-world object recognition is difficult because objects appear at many angles and take many forms.

2.2.4 Egocentric Action Recognition

Ramanan et al. [16] address the problems of object-based action recognition by leveraging an egocentric setting. The egocentric setting makes it much easier to discriminate between objects that are being interacted with and those that aren't by looking at the interactions of hands and arms with the scene. This is difficult in third-person video because hands and arms are often more difficult to identify in the scene, and they often obscure the object being interacted with. With knowledge of the objects being interacted with it is possible to ignore or weight down clutter in the rest of the scene. A separate recognition model can be learned for objects undergoing interaction, since they tend to look different to objects at rest.

Recognizing objects at different viewpoints and scales remains a difficult problem, made harder by the potential for objects to be obscured by hands or other scene objects. This is largely because existing web-based image collections like ImageNet mostly depict objects from iconic viewpoints. This leaves us with the monumental task of building an image collection containing objects depicted at many more viewpoints. If we focus the dataset specifically on activities of daily living we could narrow the object set down to those commonly found in the home - even then there's great variation within the individual object classes. This is one of the reasons why optical flow is favoured by the action recognition community; the other being that many activities are not discriminated by objects in the scene. Physical activities like running and walking are discriminated by global motions and the objects in the scene tell little.

Using optical flow alone, Kitani et al. [10] learn to recognize activities in egocentric sports videos in the absence of labelled training data. A motion histogram codebook is built and a Dirichlet process mixture model used to learn action categories. Interestingly their features contain a frequency component that captures the human gait. Whole body motion has little discriminating significance for indoor activities that are the focus of much of the research - but for sport videos it has high significance, in particular the frequency component can help to distinguish between walking and running. It seems that while indoor activities are all about local features, in particular those local to the hands, outdoor activities are all about global features. Kitani et al. use RANSAC in an attempt to completely remove local motion from the frame, in addition to noise introduced by optical flow estimation.

It's clear that the nature of the action has a high impact on the model used to recognise it. To the authors knowledge, there has been no public attempt to develop an egocentric action recognition system that works both in the indoor and outdoor environment. Likewise some state-of-the-art techniques [7] have never been applied in an egocentric setting.

2.3 Datasets

2.3.1 KTH

The KTH human action dataset is a fixed camera dataset containing 6 action types: walking, jogging, running, boxing, hand waving and hand clapping (Figure 2.3.1). Each action is performed by 25 subjects in different environments (e.g. indoor/outdoor) and clothing. The lack of variation in perspective and framing makes action recognition fairly easy, and state-of-the-art techniques achieve almost perfect performance on this dataset. It remains useful for the purpose of validating new implementations or approaches.

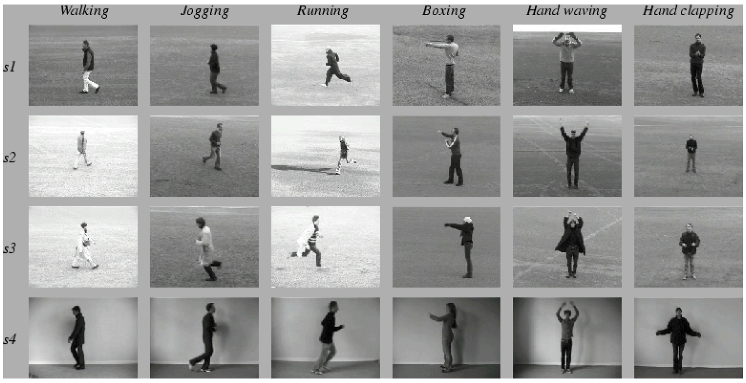


Figure 2.1: Examples from the KTH human action dataset.

2.3.2 Hollywood2

Hollywood2 is another third-person human action dataset, but it is much more realistic than the KTH dataset. 12 types of human actions are extracted from a collections of 69 movies. This dataset is challenging because there is significant variation in perspective and framing. The action types are also quite similar (e.g. hugging and kissing) which increases scope for statistical overlap. To further increase difficulty the camera frequently moves during the scene. This dataset remains a challenge but cutting edge techniques are approaching perfect performance.

2.3.3 GeorgiaTech Egocentric Activities (GTEA)

GTEA is the only well-cited human action dataset that takes place in an egocentric setting. 7 types of everyday human actions are performed by 4 subjects. This dataset was developed to test an object-based action recognition approach and as such consists of food preparation actions that are easy to recognise using an object-based approach. Specifically, the subject makes: a cheese sandwich, coffee, coffee with honey, a hotdog sandwich, a peanut-butter sandwich, a peanut butter and jam sandwich, and a sweet tea. All of these actions consist of multiple steps which increases difficulty, especially for approaches that don't use object recognition.

2.4 Types of Actions

Previous work has focused on recognizing a specific type of action. Ramanan et al. focus on multi-step actions that involve hands interacting with a number objects [16]. Kitani et al. consider outdoor sports actions [10].

In this work we're interested in recognizing actions across multiple conventional action types. Because of this, we found it helpful to develop a new taxonomy for actions.

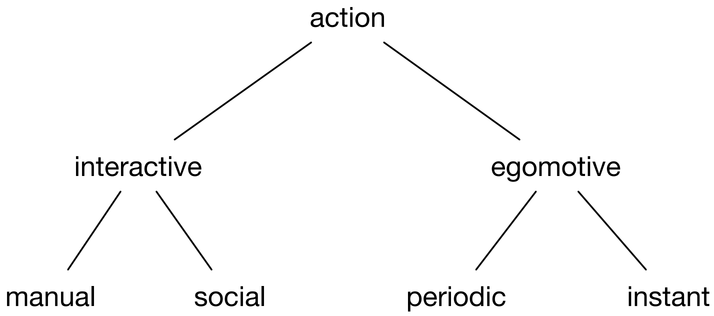


Figure 2.2: A new taxonomy for actions.

2.4.1 Definitions

In Figure 2.4, and in the rest of this work, *interactive* actions are those that involve interaction with the scene. *Egomotive* actions are those defined purely by the movement of the subject within the scene.

Examples of interactive actions are:

- making a cup of coffee
- writing a letter
- asking someone for directions
- picking up a cat

Examples of egomotive actions are:

- running
- performing a ski jump
- sitting down
- standing still

Interactive actions are further broken down into those that are *social*, e.g. having a conversation, and those that are *manual* (involving hands), e.g. solving a Rubik's Cube. Egomotive actions are broken down into those that are *periodic*, e.g. running, and those that are *instant*, e.g. sitting down. We acknowledge that this new taxonomy is incomplete (there are actions that don't fit anywhere like kicking a football), but it covers the most studied types of actions.

In this taxonomy, type divisions are informed by potential for action recognition model specialization. By which we mean, an action recognition system could specialize to each type naturally. For example to recognise manual actions the key is to understand what is happening around the hands. To recognise egomotive periodic actions it is useful to analyse the frequency components in the signal. To recognise social actions we often use face recognition.

2.4.2 Everyday Actions

This work is challenging because our goal is to recognise *everyday actions*, and everyday actions don't fall neatly into one of the above actions types. Instead, we hope to recognise actions across multiple types. Since our motivating application is automated application launching, we hope to recognise actions that might realistically be assisted by an application. This rules out instant actions, which are too short to be assisted by an application. Likewise social interactions are unlikely to be assisted by an application since applications that understand social interactions don't exist yet. This leaves us with interactive manual actions and egomotive periodic actions, both of which are well suited to hands-free assistance via action recognition.

2.5 Conclusion

To achieve our goal of recognizing everyday actions in egocentric video footage we focus on optical flow features since these are proven to be useful discriminators for many types of actions [10, 16, 7]. A new dataset must be created to evaluate our approach since none of the existing datasets cover a realistically broad set of everyday actions in an egocentric setting. Specialized approaches [10, 16] must be drawn together in order to achieve good performance on a broader dataset. Research that took place in a fixed camera setting [3] must be adapted to the egocentric setting by taking into consideration consistency in perspective and potential for model personalization.

Chapter 3

Approach

This chapter describes our approach to action recognition.

3.1 Overview

We decide to take a purely motion-based approach to the action recognition problem 3.1, for the following reasons:

- Object and object-interaction approaches are only feasible when actions contain objects and object interactions. Since we're interested in physical actions like running that are defined purely by motion signals we cannot rely upon objects or object-interactions. Object-interaction based approaches have indeed been shown to be more effective than motion-based approaches in actions that *do* contain object interactions, but for single-step everyday action recognition we submit that motion alone is adequate.
- Massive variability in the appearance of objects in the real-world means approaches that rely on object recognition require huge datasets. To illustrate this consider a packet of cereal. The variation between brands mean this object has thousands of unique appearances. Multiply this by different lighting conditions and perspectives, and then by the huge number of object classes. The result is unfathomably large, and to make matters worse the number is always growing as new new brands and designs emerge. Even if such a dataset existed, training models on such a large amount of data would itself present a significant computational challenge. Motion is invariant to much of the variation in object appearances, which allows us to use a significantly smaller dataset and maintain generality.

The first step in characterizing motion is to estimate the velocities of points in the scene.

3.2 Optical Flow

In this section, we introduce optical flow. We discuss our requirements and justify our selected optical flow algorithm.

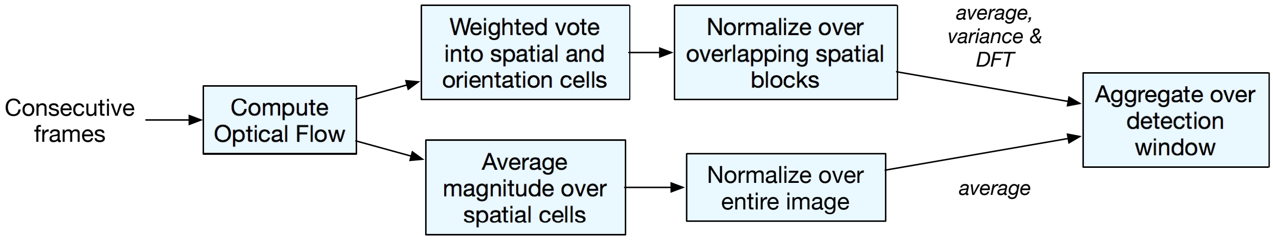

Consecutive frames
Compute Optical Flow
Weighted vote into spatial and orientation cells
Normalize over overlapping spatial blocks
average, variance & DFT
Average magnitude over spatial cells
Normalize over entire image
average
Aggregate over detection window

Figure 3.1: Feature extraction process for our action recognition approach.

3.2.1 Overview

Optical flow is an estimation of the 2D velocities between two adjacent frames. At a high-level, the idea is to find correspondences between the frames. Correspondences can be tracked at every pixel (dense optical flow) or at interest points (sparse optical flow). There are a number of common approaches to tracking correspondences:

- Gradient-based methods
- Block-matching
- Feature matching (e.g. SIFT, SURF)
- Energy-based methods

For our action recognition approach we limit our consideration to gradient-based methods, for the following reasons:

- Many of the seminal papers on optical flow are gradient-based methods [9, 13, 6].
- OpenCV includes well-tested implementations of all the most popular gradient-based methods [14].
- Gradient-based methods were used to produce state-of-the-art performance in many action recognition approaches.
- Gradient-based methods rank highly both in speed and accuracy in various benchmarks [1, 8].

3.2.2 Optical Flow Equations

The basic assumption of optical flow estimation is that pixel intensities are translated between frames, that is:

$$I(x, y, t) = I(x + u_1, y + u_2, t + dt) \tag{3.1}$$

Where $I(x, y, t)$ is the intensity of pixel (x, y) at time t, and (u_1, u_2) is the velocity of this pixel. We call this the *brightness constancy constraint*. Assuming movement is sufficiently small we can apply the Taylor series to get:

$$\frac{\delta I}{\delta x}u_1 + \frac{\delta I}{\delta y}u_2 + \frac{\delta I}{\delta t} = 0 \tag{3.2}$$

This equation has two unknowns – the brightness constancy constraint is only enough to solve for one unknown. To solve the equations for both unknowns we must introduce new constraints and different algorithms make different choices here. The most common additional constraint is the *smoothness assumption*: that motion is smooth in the local neighbourhood of a pixel.

These constraints are only true in an idealized context. For example the brightness constancy constraint does not hold when the distance of an object to the light source changes frame-to-frame, for obvious reasons. Thus instead of trying to satisfy the constraints we instead minimize the squared error in each constraint.

3.2.3 Dense vs. Sparse

In addition to selecting an algorithm we must also decide whether optical flow should be dense or sparse. Sparse optical flow has the advantage of being significantly less expensive to compute, thus if only a small region of the motion field is discriminative it is sensible to use sparse flow. In a fixed-camera setting this is often the case, for example the region containing a person. However, in an egocentric setting the entire frame is interesting. Even lack of motion in a certain region of the frame can be discriminative.

As an example we consider the handwriting action. In this case the most discriminative region of the motion field is indeed local to the writing hand, but the lack of motion in the rest of the frame is also discriminative – it represents the *motion context*.

Since motion context is discriminative for egocentric action recognition we use dense optical flow.

3.2.4 Algorithm Selection

In order to select an appropriate optical flow algorithm for our approach we evaluate several of the most popular algorithms: Horn-Schunck [9], Lucas-Kanade [13] and Farneback [6]. We use the OpenCV implementations of these algorithms since they are well tested and optimized for performance on conventional hardware. We use the default parameterisations since it is assumed these are sensible.

For our tests we use two action example: running and handwriting. Based on visual inspection we find that the Farneback algorithm produces a much less noisy flow field than either Horn-Schunck or Lucas-Kanade. This comes as little surprise since Farneback is the newer algorithm and also a top performer on the benchmarks [1, 8].

We also tested the speed of the three algorithms (Figure 3.2.4). We find that Lucas-Kanade is approximately 5 times faster than Horn-Schunck and Farneback, but Farneback is acceptably fast, coming in at approximately 50 milliseconds per frame.

Due to its superior accuracy, Farneback optical flow is used in our approach.

3.2.5 Summary

The Farneback dense optical flow is fast and provides us with a very accurate motion field. The next step is to characterize this dense motion field in a way that discriminates between action classes.

Algorithm	Running (milliseconds)	Handwriting (milliseconds)
Horn-Schunck	49	48
Lucas-Kanade	7	7
Farneback	53	47

Figure 3.2: Runtime of various optical flow algorithms per frame averaged over 100 frames. Videos have dimensions of 320x240. We use OpenCV implementations with default parameterisations.

3.3 Histograms of Optical Flow (HOF)

3.3.1 Overview

We use Histograms of Optical Flow (HOF) as our main motion descriptor. HOF descriptors are derived from Histograms of Oriented Gradients (HOG), which were originally used to produce state-of-the-art results for human recognition problems [3]. The HOG descriptor is a histogram of the intensity gradients within a rectangular region of an image. HOF is applied to the optical flow between corresponding regions in successive frames.

Formally, for a pair of images I_1 and I_2, we calculate the optical flow matrix, $F = O(I_1, I_2)$. From F we calculate the flow orientations $\theta(F)$ and magnitudes $M(F)$. We then divide the flow matrix into non-overlapping spatial cells and aggregate $\theta(F)$ into local histograms, weighted by $M(F)$. Cells are then grouped into overlapping blocks and contrast normalized to improve robustness to wide variation in local optical flow magnitude. This is essentially the same process as that outlined in [3].

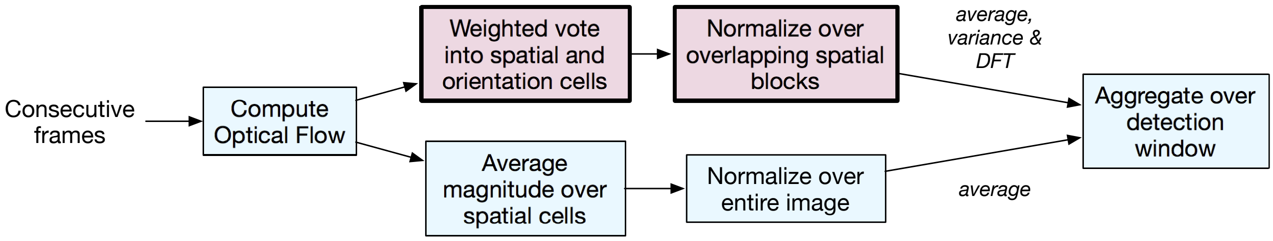

Consecutive frames
Compute Optical Flow
Weighted vote into spatial and orientation cells
Normalize over overlapping spatial blocks
average, variance & DFT
Aggregate over detection window
Average magnitude over spatial cells
Normalize over entire image
average

Figure 3.3: Highlighting the HOF step in the feature extraction process.

3.3.2 Mapping Orientations to Histogram Bins

Flow orientations are traditionally mapped to histogram bins in a way that makes the HOF descriptors directionally invariance [3] (Figure 3.3.2). For example, a right-hand-wave produces the same descriptors as a left-hand-wave. This is often useful in the context of a fixed-camera setting, in which directionality is rarely a distinguishing quality. However in a personalized egocentric setting directionality is often a useful cue. A right-handed subject engaged in the hand writing action will produce an oscillatory motion with different directionality to that of a left-handed subject. By disabling directional invariance we enable personalized learning.

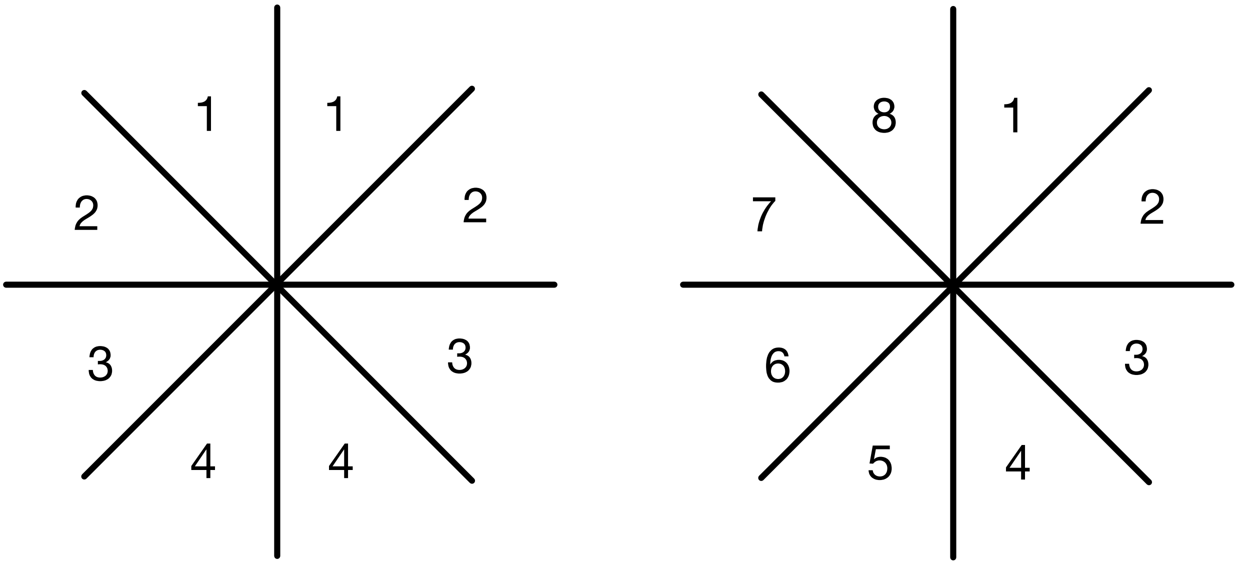


Figure 3.4: Different ways of mapping orientations to 8 bins. Left: Directionally invariant in the horizontal axis - that is, a right-hand-wave produces the same descriptors as a left-hand-wave. Right: Directionally variant, a right-hand-wave produces different descriptors to a left-hand-wave.

3.3.3 Global vs. Local

Our action recognition approach is based on global features densely sampled across the entire frame. This is counter to the norm, HOF/HOG and other popular image or video descriptors are usually computed at interest points, or densely within some region of interest [3], for the following reasons:

- Descriptor computation is often slow and sparse sampling decreases the number of descriptors.
- In a conventional fixed-camera setting we're usually only interested in a small region of the image, for example the region containing a human. Features computed outside of this region are of little use to learning and will probably hinder it.
- We sometimes know in advance that certain objects are very informative. For example in certain action recognition problems it is useful to detect hands first and compute descriptors in close proximity to the hands.

The global locations of these descriptors are usually omitted from the models representation of the scene. This is because in many fixed-camera recognition problems, objects or events in the scene mean the same thing regardless of location in the frame. For example, a ball is still a ball wherever it appears in the frame. By omitting the global locations of the descriptors the model becomes positionally invariant.

The egocentric setting is quite different. Since the camera perspective is consistent with respect to the subject, descriptor location becomes discriminative. For example, a right-handed subject engaged in the handwriting action can be relied upon to produce the flow pattern that corresponds to handwriting in the bottom-right of the frame. Conversely, if this flow pattern is observed elsewhere in the frame we can be quite sure that it doesn't correspond to handwriting. For this reason, in an egocentric setting global features are appropriate.

3.3.4 Local Contrast Normalization

Dalal et al. showed that local contrast normalization has a significant effect on recognition performance [2, 3]. By grouping cells into overlapping blocks for contrast normalization they improve recognition accuracy by around 5%. The reasoning for this is that overlap means each HOF descriptor contributes to more than one classification feature. We use overlapping blocks with L2-normalization.

3.3.5 Cell and Block Size

Through trial and error we find that good classification performance is achieved with a cell size of 16x16 pixels combined with a block size of 2x2 cells and a block overlap of 1 cell. For 320x240 video this produces 300 blocks per detection window. We use 8 orientation bins, which means the total feature vector length for the HOF features is $300 \times 8 = 2400$.

3.4 Detection Window

3.4.1 Overview

In order to reduce noise we aggregate features over a sliding detection window of 60 frames (2 seconds). This means that action classification of a given frame depends on the 30 frames prior and after. We experimented with a gaussian weighting of the detection window based on the idea that features become less relevant the further they are from the current frame, but this had little effect on classification accuracy.

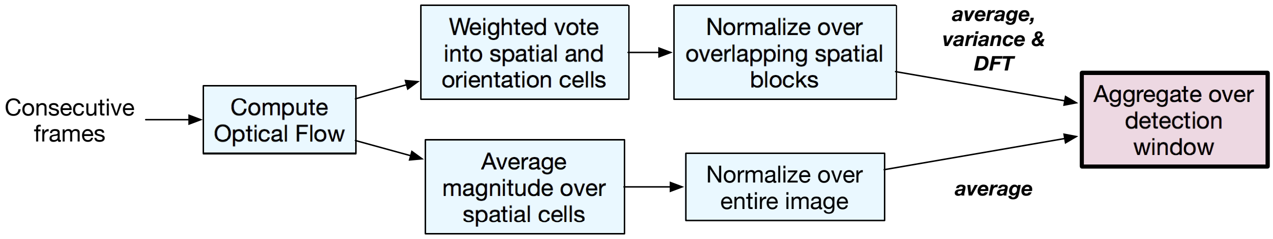

Consecutive frames
Compute Optical Flow
Weighted vote into spatial and orientation cells
Normalize over overlapping spatial blocks
average, variance & DFT
Average magnitude over spatial cells
Normalize over entire image
average
Aggregate over detection window

Figure 3.5: Highlighting the feature aggregation step in our approach.

3.4.2 HOF Aggregation

HOF features are aggregated over the detection window by both their average and variance over time. This doubles the dimensionality from 2400 to 4800.

3.4.3 Periodic Vertical Motion

One of the problems with using HOF features alone is that activities like walking and running produce feature vectors with significant overlap. This is because the aggregate motion is basically identical:

- Flow emanates from a focus of expansion which corresponds to the subjects direction of motion.
- Flow exhibits high variance in the vertical component which corresponds to the subjects step.

Thus to make periodic actions like this separable we introduce a feature that characterizes the frequency component of motion. This is achieved by computing the discrete Fourier transform (DFT) of the vertical component of the HOF features over the detection window. This idea came from [10] where DFT coefficients are used to characterize periodic motion in extreme sports videos. The top 5 DFT coefficients are concatenated to the final feature vector.

3.4.4 Magnitude

When testing our approach we found that inclusion of a globally normalized magnitude image improved classification performance. Optical flow magnitudes are averaged within 16x16 pixel non-overlapping cells and then the averages are L2-normalized over the entire image. We average the result over the detection window to produce 300 features which are concatenated to the final feature vector.

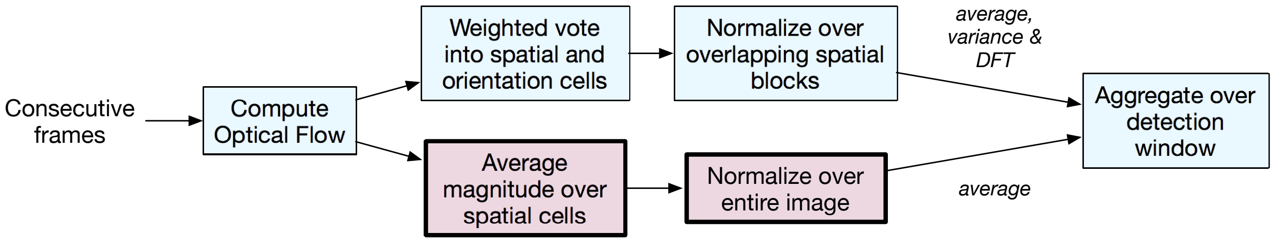

Consecutive frames
Compute Optical Flow
Weighted vote into spatial and orientation cells
Normalize over overlapping spatial blocks
average, variance & DFT
Average magnitude over spatial cells
Normalize over entire image
average
Aggregate over detection window

Figure 3.6: Highlighting the magnitude capture step in the feature extraction process.

3.5 Classification Model

3.5.1 Feature Vector

Our final feature vector has dimensionality 5105. This high dimensionality is handled well by the popular Support Vector Machine (SVM) classifier.

3.5.2 Multiclass SVM

SVMs are binary classifiers. Since our approach must be capable of recognizing more than two action we must adapt the SVM to work with multiple classes. A common way to do this is to learn a separate SVM for each pair of classes. A test case is then classified using all the classifiers and each votes for the final prediction using its binary output. The class with the most votes at the end of this process is the final prediction.

3.6 Conclusion

Our approach utilizes the popular HOF motion descriptor among other motion-based descriptors. We adapt approaches previously successful in a fixed camera setting for the egocentric setting. These adaptions mean our feature extraction process is largely custom and as such will require custom software. The implementation should make as much use of existing code as is possible.

Chapter 4

Implementation

In this chapter we discuss implementation of software to realize the approach described in the previous chapter.

4.1 Overview

The software for this action recognition approach was written in Python. Open source code was leveraged where possible, and much of the code developed for this research has been released as an open source Python package called ActiPy[1].

4.2 Technologies Used

- **OpenCV2** OpenCV2 was used for optical flow estimation. The library contains implementations of the most popular algorithms including Horn-Schnuck, Lucas-Kanade and Farneback. Of particular interest is the Farneback code which can optionally leverage the GPU for acceleration. The GPU code is not used in this work since we lacked access to GPU hardware.
- **SciPy** SciPy contains many popular machine learning algorithms, of which we use the multi-class SVM. It also contains a number of useful statistical functions which we make significant use of.

4.3 ActiPy

Our ActiPy library is designed primarily to make it easy to iterate through windows of optical flow and aggregate across them in interesting ways. A brief overview of the important software modules and classes is given below:

- `actipy.optical_flow.OpticalFlow` wraps optical flow algorithms provided by OpenCV 1 & 2 under a common API so they can be swapped in and out easily for evaluation.
- `actipy.optical_flow.Flow` represents the optical flow between two adjacent frames and contains useful functions for visualizing the flow.

[1] https://github.com/lebek/ActiPy

- `actipy.optical_flow_features.OpticalFlowFeatures` extracts features from optical flow e.g. HOF and magnitude.
- `actipy.video_features.VideoFeatures` extracts features from sequences of optical flow.
- `actipy.plan` contains utility functions that help to find good parameterisations for the feature extractors, for example `good_cells()` which finds a grid size that divises the video dimensions without remainder.
- `actipy.train` trains a model for action recognition and saves it to file.
- `actipy.hist_plot` outputs a novel visualization of HOF features.
- `actipy.dissertation` loads a trained model and evaluates it.

Chapter 5

Evaluation

In this chapter we describe a new dataset for action recognition and evaluate the performance of our approach.

5.1 Overview

Our goal is to recognise everyday actions from egocentric video. We hope to prove our hypothesis that this can be achieved using motion-based feature alone. We evaluate our approach using the dataset described in the previous section.

5.2 Dataset

5.2.1 Requirements

We set out to build a dataset consisting of labelled action videos recorded from an egocentric perspective. Our initial goal was to build a practical action recognition approach suitable for tasks such as patient monitoring and mobile worker tracking. With this in mind, it was important for the dataset to include a broad range of everyday actions. We were particularly interested in capturing both low-intensity indoor actions such as working at a computer or writing, and high-intensity outdoor actions such as running. We imposed the following additional requirements:

- **Size** based experience with other action recognition datasets we predicted that at least 10 minutes of total footage per action category would be necessary to learn a model that could generalize.
- **Variation** within each action category we aimed to ensure variation in location and lighting conditions.
- **"Unknown" category** having an "Unknown" category enables us to learn a model that can handle unseen action categories gracefully.
- **Fixed length** each example in the dataset should be cropped to contain a single action and every example should have the same length. The purpose of this is to prevent the ability to predict the category correctly based solely on the video length.

5.2.2 Design

We settled on 5 actions: walking, running, handwriting, idling, and “unknown”. We record 15 minutes of footage for each category across 3 different locations. Actions were performed naturally and then cropped after the fact into 2 second segments. The “unknown” category contains actions like opening and closing doors, making a bed, brushing teeth, and filling a cup of water.

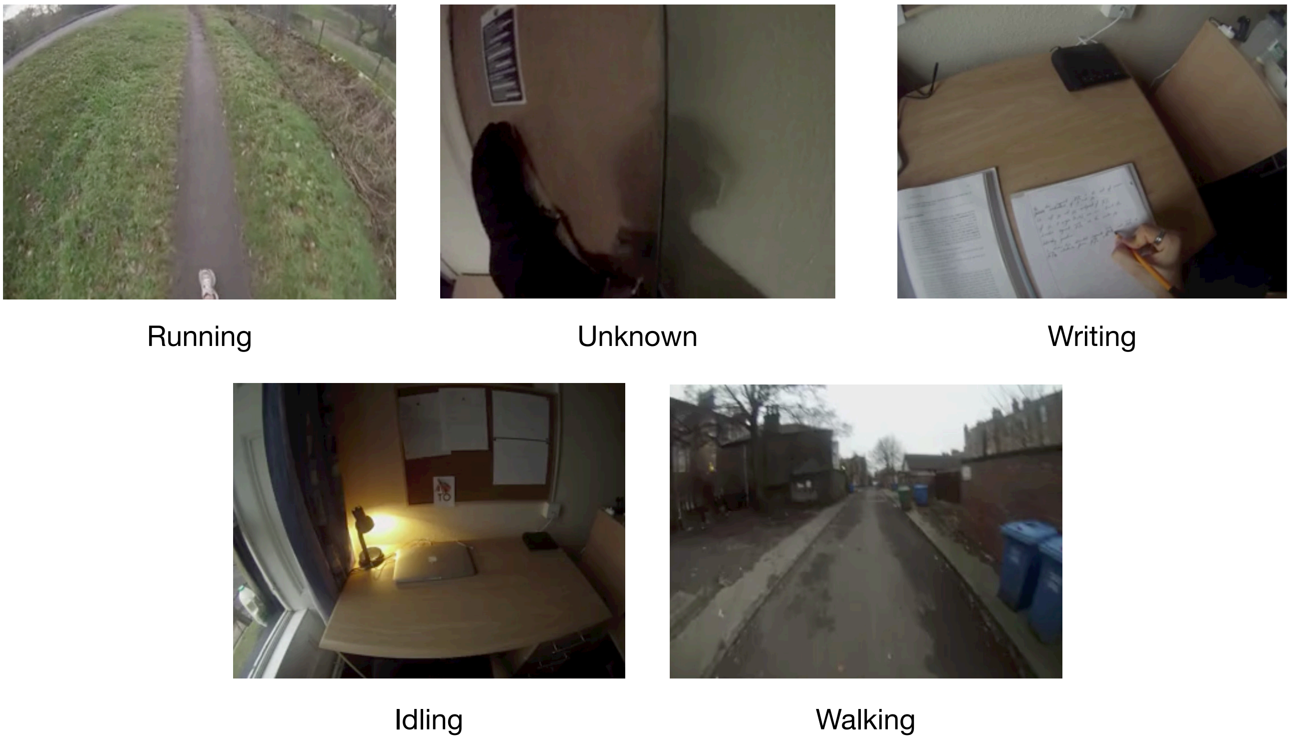
5.2.3 Collection

The dataset was collected using a GoPro HERO3+ head-mounted camera 5.1. The head-strap accessory allows the camera to be directed downward enough to capture hands and arms even when looking toward the horizon.

Figure 5.1: GoPro HERO3+ camera with head-strap.

5.2.4 Result

Some examples from each of the classes in the resulting dataset are shown in Figure 5.2.



Running

Unknown

Writing

Idling

Walking

Figure 5.2: Examples from our egocentric actions dataset.

5.3 Validation

5.3.1 Overview

Test data was used to validate the correctness of our approach and implementation. In general what we look for is evidence that our features: (a) represent what we expect them to represent; (b) appear discriminative between different action classes.

5.3.2 Optical Flow & HOF

We visually inspect Optical Flow and HOF feature quality for many action examples, three of which are shown in Figure 5.3. We find that Farneback Optical Flow is very accurate and noise free for both interactive actions (writing) and egomotive actions (running, walking). Flow is noticeably noisy in dark conditions and in the presence of motion blur, but this is a limitation of all popular Optical Flow algorithms. In our dataset motion blur is only problematic during rapid head rotation which happens infrequently. The slight motion blur and vibration present in running examples doesn't seem to be a problem for the Farneback algorithm.

Any noise present in the raw optical flow is invisible after HOF feature extraction and aggregation over the 60 frame detection window. HOF features extracted for walking and running examples are distinguishable from other actions by their emanation of flow from a focus of expansion. Writing examples contain a blob of high-variance almost-random motion in the lower right of the frame (the subject was right handed).

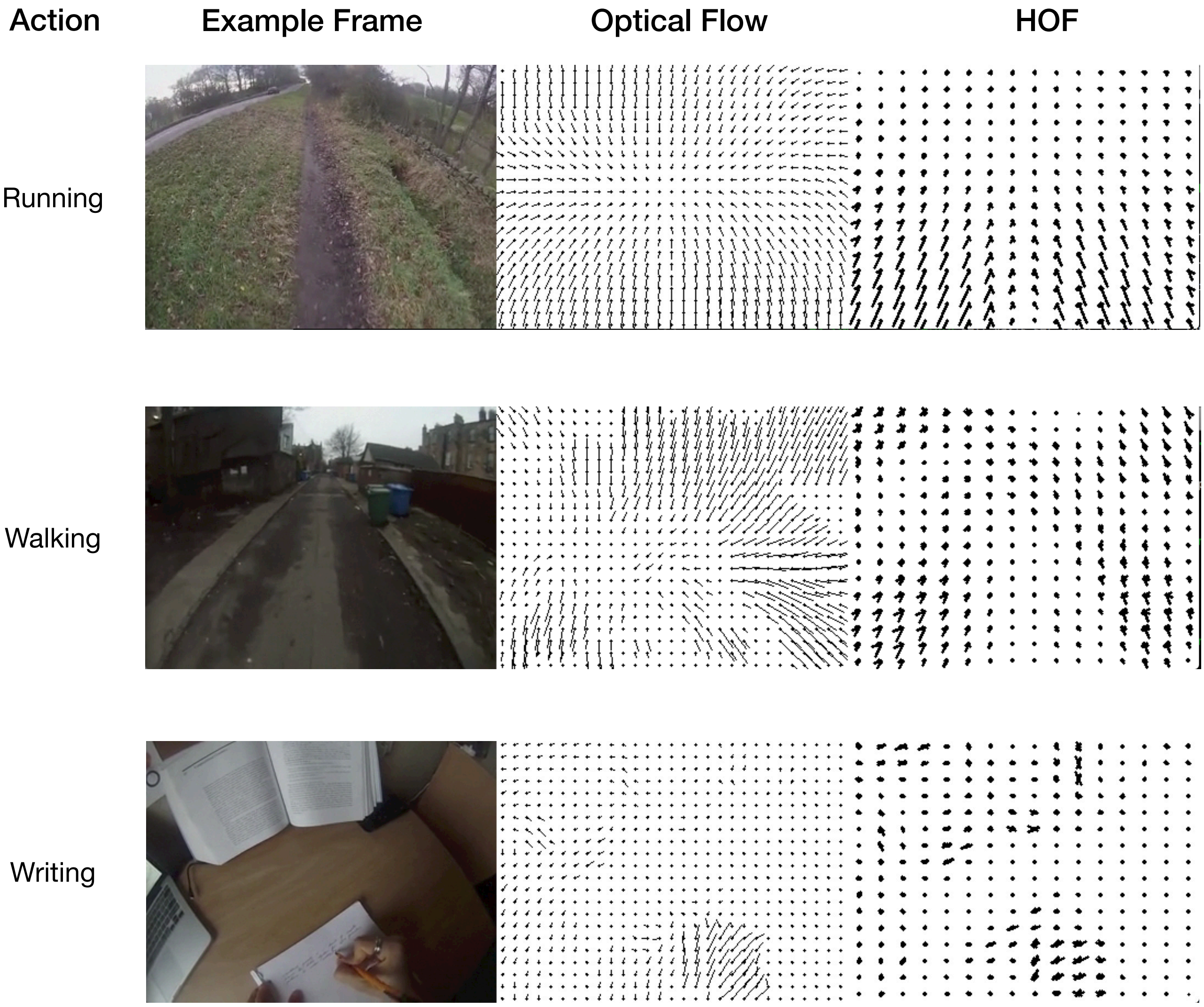


Figure 5.3: Examples of Optical Flow and HOF features computed for different action examples. We show Optical Flow for the given frame, and HOF features extracted over a window of 60 frames centered on the given frame.

5.3.3 Magnitude and Variance

Examples of magnitude and variance images extracted for some example videos are shown in Figure 5.4. Magnitude images for running and writing are insightful. In the writing example we see a blob of high motion where the hand is. In running we see regions of high motion emanate from the focus of expansion.

Variance images are less easy to interpret, but at least in the running video we can see that variance is greatest at the focus of expansion. We expect this is because it represents the deepest point in the scene and estimated flow for faraway objects is noisy.

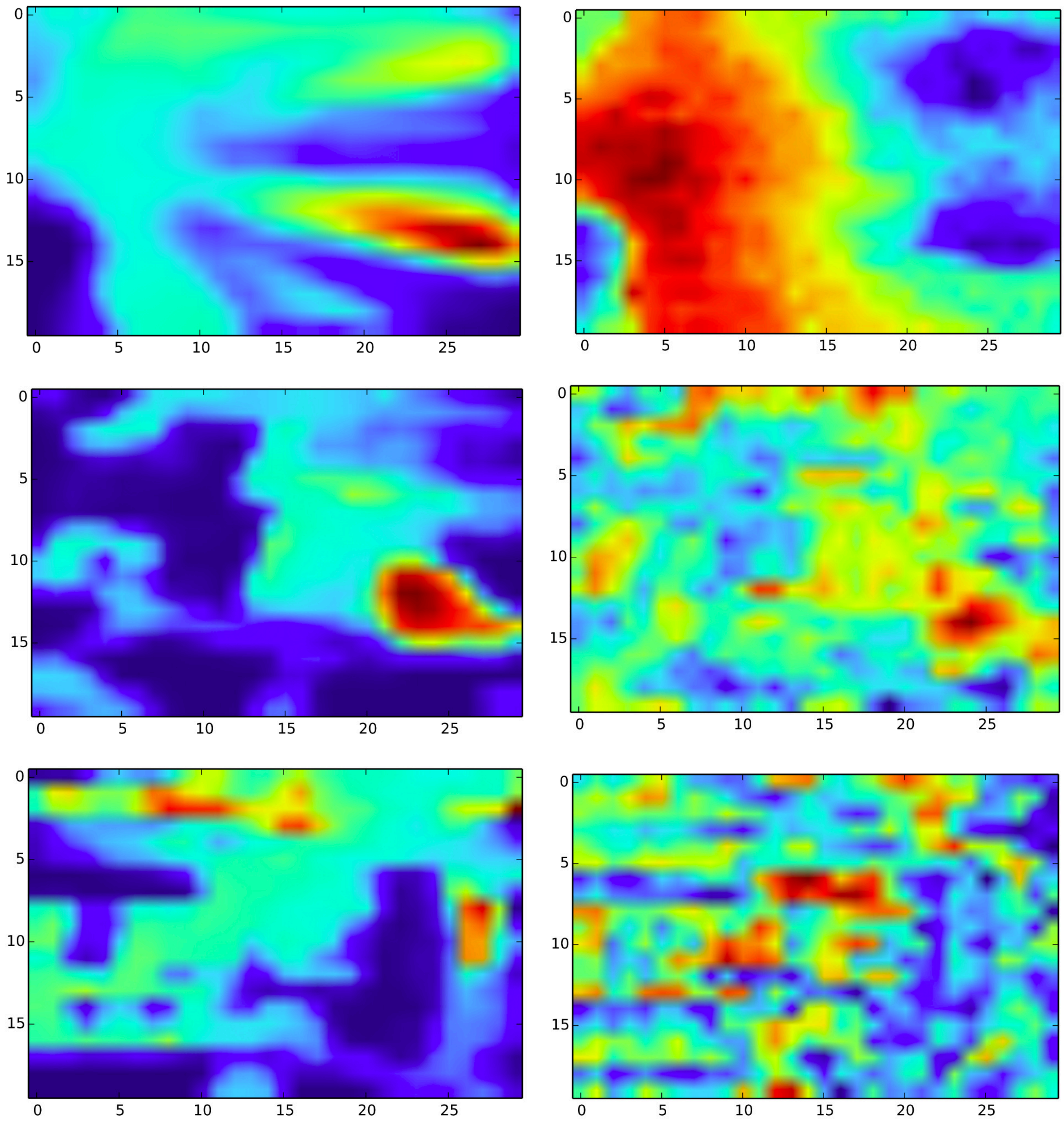


Figure 5.4: Analysis of variance and magnitude in action videos. Top to bottom: running, writing, idling. Left: magnitude, right: variance.

5.3.4 Periodic Motion

Periodic activities with similar motion are differentiated by analysing the frequency component of motion. We do this by computing the discrete Fourier transform (Figure 5.5) and taking the top 5 coefficients.

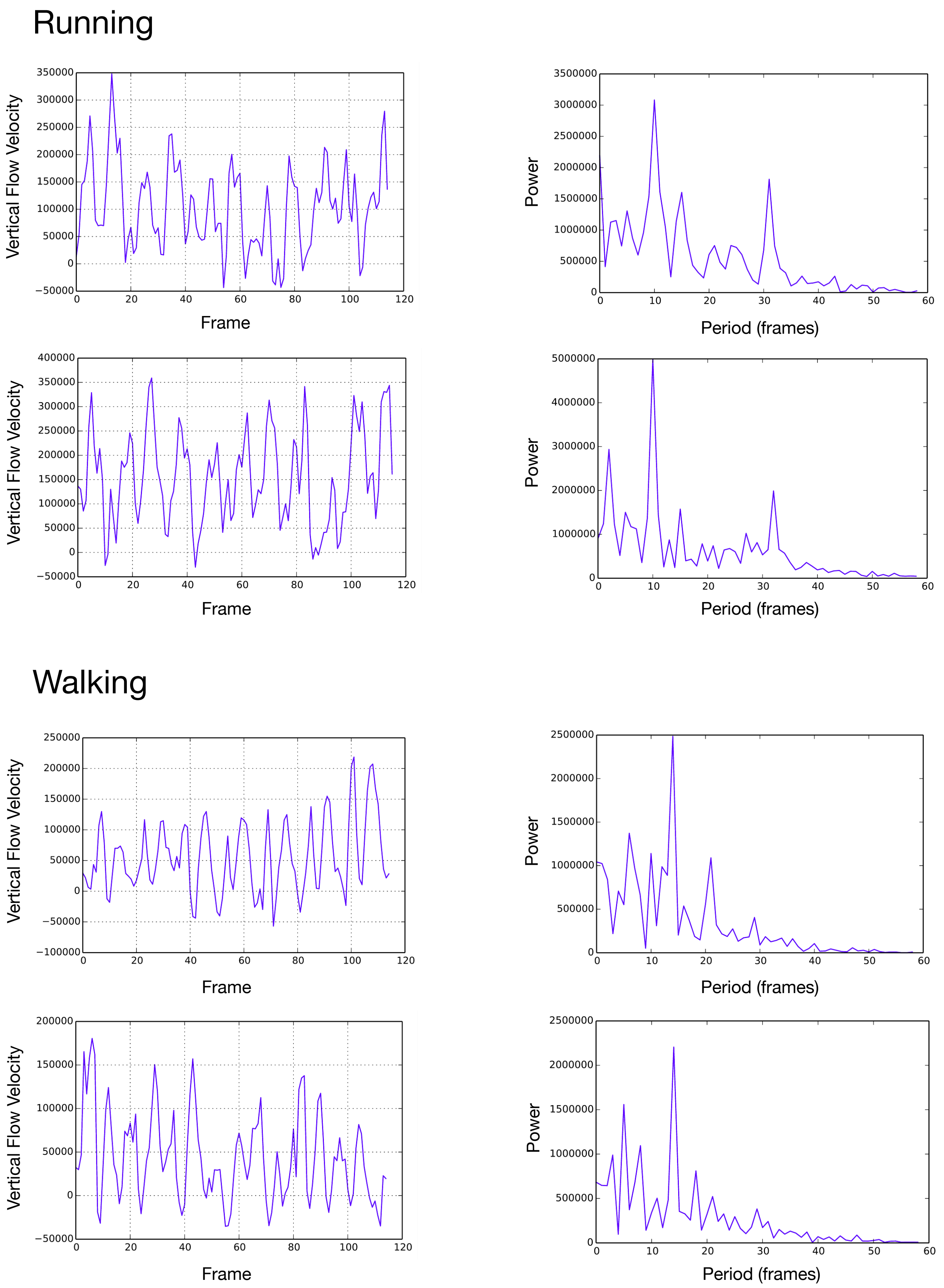


Figure 5.5: Analysis of periodic motion in running and walking videos. Graphs on the left show the net vertical flow velocity over time for two running and walking examples. Periodograms on the right show the corresponding peak DFT coefficients.

Running has DFT peaks at 10 and 30 frames which the author speculates corresponds to step and breathing rate respectively. Video in our dataset is recorded at 30 fps, so this would correspond to a step rate of 3 per second, which is consistent with runner advice to run 180 steps per minute [15]. Walking has peaks at around 14 frames, making it about 50% slower than running.

5.4 Experimental Setup

Our test set consists of 6 actions performed for 15 minutes each in different environments. Each example is broken down into 2 second segments, which are then downsized to 320x240 pixels. We extract features from each segment as described in our approach and use them to train a multiclass SVM.

Our test set consists of 3 continuous videos in which the 6 actions are performed in succession for roughly the same amount of time each (Figure 5.5.1). An example of one of these videos is given in Figure 5.6.

Test videos are labelled by hand at frame resolution. We recognise that there is some degree of subjectivity to this. For example it is not clear when switching from running to walking the exact frame at which the switch occurs. In general, for each action transition, we consider the range of possible opinions and pick the median.

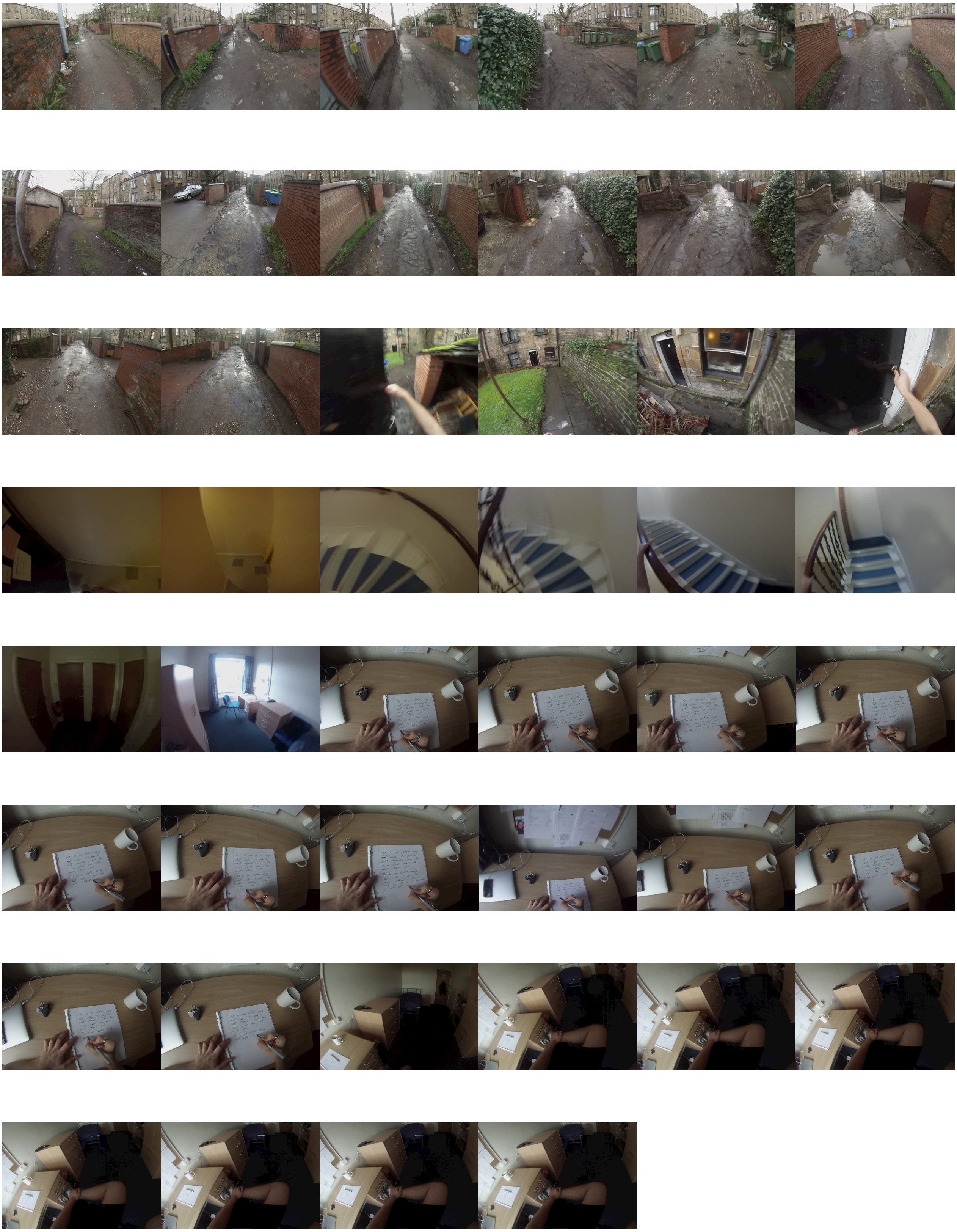


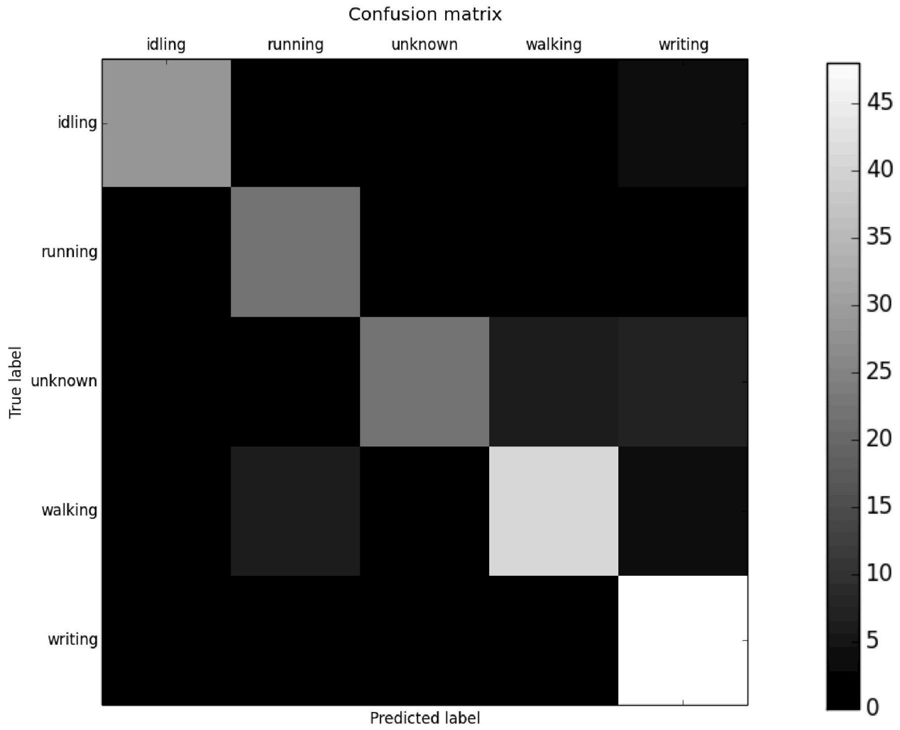
Figure 5.6: Example test video containing all 5 actions.

5.5 Performance

5.5.1 Overview

Figure 5.5.1 demonstrates the prediction accuracy. Our approach recognizes actions with an impressive overall classification accuracy of 82.67%. Unknown actions are generally classified as such which is particularly impressive because it means the model can separate what it can classify from what it can't. This significantly reduces the number of false positives in real world scenarios where unknown actions are likely to be the most common by a significant margin.

We see that walking is sometimes confused with running which upon further investigation occurs when the subject is walking faster than usual. A larger dataset with more examples of both walking and running would help to improve the separation of these classes, in-particular in dimensions that correspond to periodic motion features.



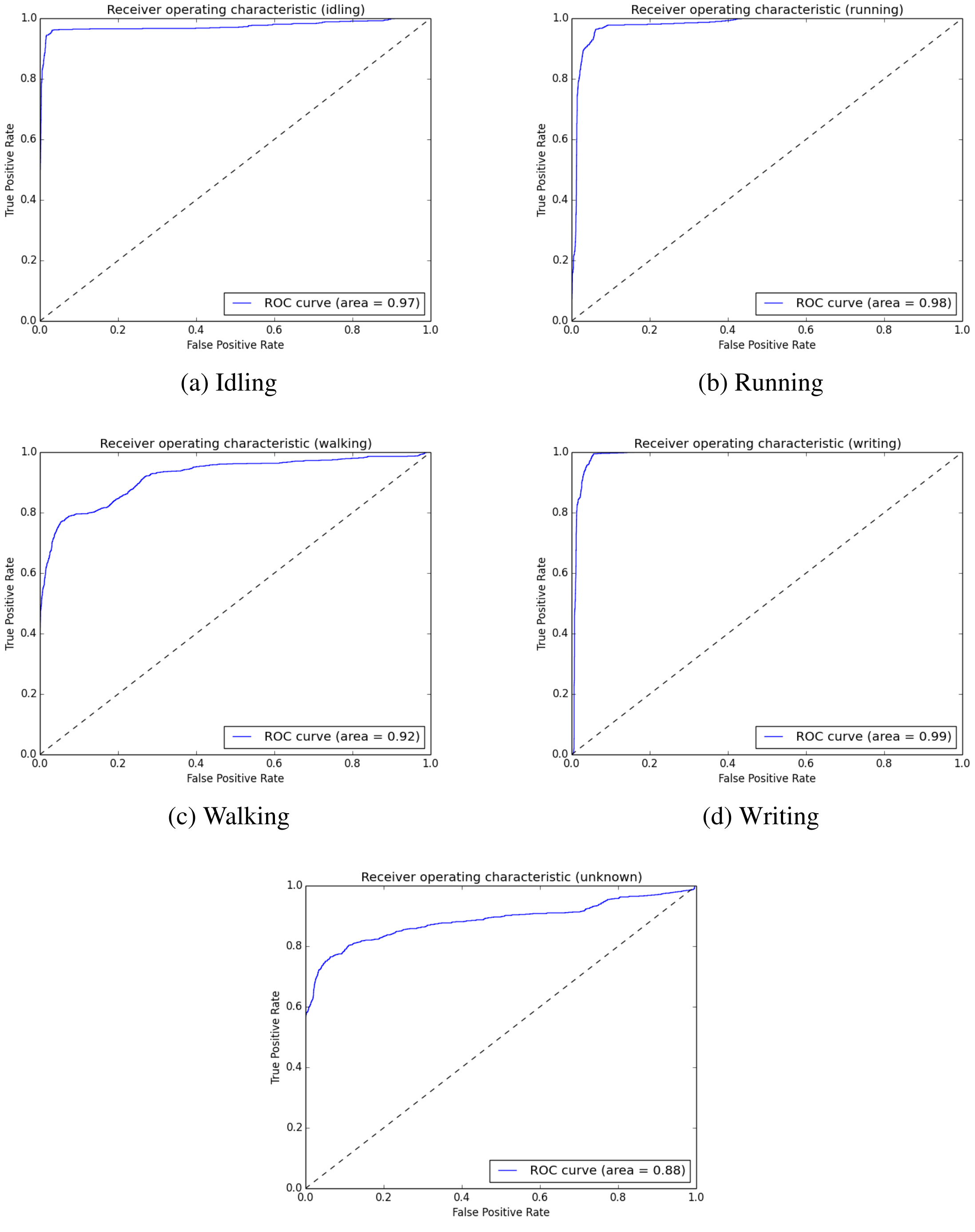
Test data	Duration (mins:secs)	Accuracy
Test video 1	3:00	77.8%
Test video 2	3:19	87.4%
Test video 3	4:02	83.9%
Test video 4	3:26	80.9%
All videos	13:47	82.67%

Figure 5.7: Above: Confusion matrix for action recognition on the full test data. Below: Break down of duration and accuracy scores for individual test sequences.

5.5.2 False Positives

Our motivating application is automated application launching on head-mounted devices. In this application false positives for actions are highly undesirable because they result in an application launching for the wrong action. For example the running application launches as the subject sits down. Thus while we can tolerate false positives for the unknown action – that is, for example, running is classified as unknown and no application launches – we cannot tolerate false positives for anything else. A ROC analysis (Figure 5.5.2) demonstrates the success of our approach in this respect.

Idling, running and writing examples are classified with almost perfect accuracy, with walking examples following closely. The unknown class, on the other hand, must take on a false positive rate of 20% to achieve an 80%+ true positive rate. This is not a serious problem since occasional false positives for the unknown class are acceptable.



(a) Idling

(b) Running

(c) Walking

(d) Writing

(e) Unknown

Figure 5.8: Receiver operator characteristic curves for action recognition on the full test data.

Chapter 6

Conclusion

6.1 Summary

In this work our goal was to recognise everyday actions in egocentric video footage. We were able to successfully adapt state-of-the-art motion-based descriptors to the egocentric setting, achieving a prediction accuracy of 82.67%. This supports our hypothesis that motion-based descriptors are enough for action recognition, which greatly reduces the size of the dataset from that which would be required if using an object-based approach.

We adapt conventional motion descriptors to the egocentric setting by leveraging the consistent perspective. We find that performance can be improved by supplementing conventional features with features designed for the egocentric setting. In our case, performance was improved by integrating a new motion frequency-based feature.

Our original motivation was automatic application launching on head-mounted devices. Based on our evaluation we believe the approach described in this work is fit for purpose.

6.2 Future Work

We demonstrate that actions can be recognized in egocentric video using motion-based features alone for a set of 5 everyday actions. As immediate future work we propose this approach be evaluated on a larger set of actions. State-of-the-art approaches should also be evaluated against the expanded dataset for comparison.

6.2.1 Higher-level Action Features

This work is primarily focused on finding the right representations of motion at a low-level of abstraction. As a result, the treatment of higher-level action features like transitions, multi-step actions and background actions is fairly rudimentary. This leaves much scope for future work.

Our approach models actions, but not the transitions between them. This property results in poor recognition performance during the transition. Transitions cover a relatively short timespan compared to the actions themselves, and often contain intense, unpredictable, motion. This raises questions as to how transitions should be modelled and whether they require a separate model to the actions themselves. Future work could develop models for transitions between actions and evaluate the benefits of learning such models.

Another interesting avenue for future work is modelling the probability of transitioning between specific actions. For example it is more likely to transition from running to walking than from running to writing. Our approach doesn't model state change probabilities in any way.

The fixed-width time window used in our approach is inappropriate for recognizing multi-step actions since: (a) it can't represent a collection of separate steps; (b) it can't accurately capture steps that are shorter than the time window. A more sophisticated treatment of time, perhaps following the actom model set out by Gaidon et al. [7], would enable recognition of multi-step actions. The new approach could explore allowing actions and action-steps to overlap, thus allowing background actions.

6.2.2 Personalization

Due to time constraints our dataset was limited to a single subject, which meant we weren't able to evaluate the effects model personalization had on performance. As future work we propose the dataset be expanded to include several subjects performing the same set of actions. The effects of model personalization could then be evaluated empirically.

For model personalization to be useful in practice we need to consider how we can learn a new model for each subject without requiring each subject to arduously collect data before using the system. The author proposes a hybrid approach whereby a generic model is personalized over time. That is, we first learn a generic model by using a dataset consisting of many subjects. We incrementally personalize the generic model for each user by doing online learning. This approach would of course depend on the generic model working quite well already, since it would be used to assist the personalized updates. Future work focused on this problem would bring egocentric action recognition closer to commercial application.

6.2.3 Other Sensors

Egocentric vision research is becoming a more popular as consumer head-mounted camera devices get closer to market. These devices will likely develop to include accelerometers and gaze trackers. Egocentric action recognition research should evolve to utilize these sensors in addition to the video feed. As an example, the accelerometer could be used to remove background motion from the optical flow, which is often desirable. Gaze tracking provides useful attentional cues.

Bibliography

[1] Simon Baker et al. "A Database and Evaluation Methodology for Optical Flow". English. In: *International Journal of Computer Vision* 92.1 (2011), pp. 1–31. ISSN: 0920-5691. DOI: 10.1007/s11263-010-0390-2. URL: http://vision.middlebury.edu/flow/.

[2] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection". In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE. 2005, pp. 886–893.

[3] Navneet Dalal, Bill Triggs, and Cordelia Schmid. "Human detection using oriented histograms of flow and appearance". In: *In European Conference on Computer Vision*. Springer, 2006.

[4] O. Duchenne et al. "Automatic annotation of human actions in video". In: *Computer Vision, 2009 IEEE 12th International Conference on*. 2009, pp. 1491–1498. DOI: 10.1109/ICCV.2009.5459279.

[5] Alexei A. Efros et al. "Recognizing Action at a Distance". In: *IEEE International Conference on Computer Vision*. Nice, France, 2003, pp. 726–733.

[6] Gunnar Farnebäck. "Two-frame motion estimation based on polynomial expansion". In: *Image Analysis*. Springer, 2003, pp. 363–370.

[7] Adrien Gaidon, Zad Harchaoui, and Cordelia Schmid. "Actom sequence models for efficient action detection." In: *CVPR*. IEEE, 2011, pp. 3201–3208. URL: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2011.html#GaidonHS11.

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012. URL: http://www.cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=flow.

[9] Berthold K Horn and Brian G Schunck. "Determining optical flow". In: *1981 Technical Symposium East*. International Society for Optics and Photonics. 1981, pp. 319–331.

[10] K.M. Kitani et al. "Fast unsupervised ego-action learning for first-person sports videos". In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. 2011, pp. 3241–3248. DOI: 10.1109/CVPR.2011.5995406.

[11] I. Laptev et al. "Learning realistic human actions from movies". In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587756.

[12] Q.V. Le et al. "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis". In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. 2011, pp. 3361–3368. DOI: 10.1109/CVPR.2011.5995496.

[13] Bruce D Lucas, Takeo Kanade, et al. "An iterative image registration technique with an application to stereo vision." In: *IJCAI*. Vol. 81. 1981, pp. 674–679.

[14] "Motion Analysis and Object Tracking". In: (). URL: http://docs.opencv.org/modules/video/doc/motion_analysis_and_object_tracking.html.

[15] Kelly O'Mara. "Make A High Stride Rate Work For You". In: URL: http://running.competitor.com/2013/08/training/make-a-high-stride-rate-work-for-you_54957.

[16] Deva Ramanan. "Detecting Activities of Daily Living in First-person Camera Views". In: *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. CVPR '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 2847–2854. ISBN: 978-1-4673-1226-4. URL: http://dl.acm.org/citation.cfm?id=2354409.2355089.

[17] Scott Satkin and Martial Hebert. "Modeling the Temporal Extent of Actions". In: *Computer Vision ECCV 2010*. Ed. by Kostas Daniilidis, Petros Maragos, and Nikos Paragios. Vol. 6311. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, pp. 536–548. ISBN: 978-3-642-15548-2. DOI: 10.1007/978-3-642-15549-9_39. URL: http://dx.doi.org/10.1007/978-3-642-15549-9_39.

[18] Scott Satkin and Martial Hebert. "Modeling the Temporal Extent of Actions". In: *Computer Vision ECCV 2010*. Ed. by Kostas Daniilidis, Petros Maragos, and Nikos Paragios. Vol. 6311. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, pp. 536–548. ISBN: 978-3-642-15548-2. DOI: 10.1007/978-3-642-15549-9_39. URL: http://dx.doi.org/10.1007/978-3-642-15549-9_39.

[19] C. Schuldt, I. Laptev, and B. Caputo. "Recognizing human actions: a local SVM approach". In: *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. Vol. 3. 2004, 32–36 Vol.3. DOI: 10.1109/ICPR.2004.1334462.

[20] Heng Wang et al. "Evaluation of local spatio-temporal features for action recognition". In: *University of Central Florida, U.S.A.* 2009.