Spatio-Temporal Optical Flow Statistics (STOFS) for Activity Classification

Vignesh Jagadeesh*
Dept. of ECE
University of California
Santa Barbara, CA-93106
vignesh@ece.ucsb.edu

S. Karthikeyan†
Dept. of ECE
University of California
Santa Barbara, CA-93106
karthikeyan@umail.ucsb.edu

B.S. Manjunath
Dept. of ECE
University of California
Santa Barbara, CA-93106
manj@ece.ucsb.edu

ABSTRACT

This paper presents a novel descriptor for activity classification. The intuition behind the descriptor is "learning" statistics of optical flow histograms (as opposed to learning "raw" histograms). Towards this end, an activity descriptor capturing histogram statistics is constructed. Further, a technique to make the feature descriptor scale-invariant and parts-based is proposed. The approach is validated on a dataset collected from a camera network. The data presents a challenging real world scenario (variable frame rate recording, significant depth disparity, and severe clutter), where biking, skateboarding, and walking are activities to be classified. Experimental results point to the promise of the proposed descriptor in comparison to state of the art.

Keywords

Motion Descriptor, Action Classification, Flow Statistics

1. INTRODUCTION

Activity inference from video has evoked considerable interest in the vision community for quite some time. Significant advances in modeling and inference techniques over the past decade have greatly enhanced state of the art methods leading to a renewed interest. Applications of activity inference techniques can be far reaching. For example, analyzing mobility patterns in camera networks, anomaly detection in surveillance feeds, and gesture interpretation for robotics rely critically on the underlying methods for recognizing actions. Almost all methods attempting to solve the problem adopt a two stage approach, namely low-level feature extraction and feature classification. This work follows suit, however laying greater emphasis on low-level feature representations. The application motivating the ensuing discussions is the analysis of activity in a camera network situated inside a university campus. Though the ultimate goal would be fusing activity information across multiple cameras, the current work attempts to solve the problem in a single camera setting.

This work is inspired by Efros et al's 30-pixel man [5], and Roth and Black's idea [9] of learning optical flow statistics for natural image sequences. The relation to medium resolution activity recognition [5] stems from the following scenario. Consider a camera looking at a scene with reasonable amount of depth disparity. One can easily observe an object starting as a 3-pixel man, turning into a 30-pixel and 300-pixel man (the reverse is also true). The activity descriptor for this scenario must be invariant to massive scale changes, or the inference technique must be designed to be immune to possible noise due to scale changes.

Roth and Black's idea [9] of learning flow statistics from natural image sequences could be adopted to learn flow statistics of activities. Specifically, this work concerns itself with learning Spatio-Temporal Optical Flow histogram Statistics(STOFS). It should be clarified that some recent work on optical flow histograms exists [2, 3], however the contribution of this paper lies in learning statistics of the same for activity inference. An interesting way of looking at STOFS is as a metric for capturing deformotion (deformation + motion) in video sequences.

The main contributions of this paper are:

- Novel activity descriptor (STOFS) to learn statistics of optical flow histograms
- Scale Invariant and Parts-Based STOFS
- A new dataset for 3-way classification between pedestrians/skateboarders/bikers (created from a camera network) with over 7000 frames being manually annotated for ground truth.

2. RELATED LITERATURE

Contemporary activity classification from video can be categorized from various points of view, the two most important being *sparse/dense methods* and *supervised/unsupervised methods*. Sparse activity analysis techniques detect interesting patches in the spatio-temporal volume, thereby ignoring regions that the detector deems "uninteresting". Dollar et al's sparse spatio-temporal features [4] and Laptev et al's

*Corresponding author
†Corresponding author

ICVGIP '10, December 12-15, 2010, Chennai, India

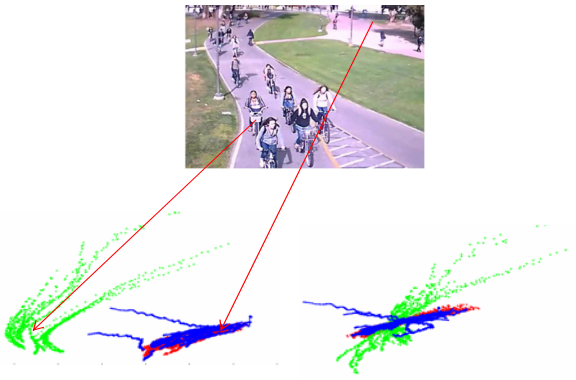

Figure 1: *Examples of activity trajectories extracted from the video. Green Trajectories represent bikers, red trajectories represent skateboarders, and blue trajectories represent pedestrians. Observe how the trajectories must be mean normalized (bottom right) to get rid of directionality that might bias the algorithm from generalizing to other data.*

STIP features [6] are state of the art space-time interest point detectors. Following the detection process, three dimensional patches are extracted around the detected interest points, and described by a feature vector (*eg: HOG, SIFT, FLOW etc..*). On the other hand, dense methods take the data as such, without sampling for interest points and attempt designing classification methods for a large number of features (extracted possibly with outliers).
Supervised methods [6, 4], (as the name implies) explicitly furnish training data to the algorithm, usually in the form of ground truth annotations. Unsupervised methods [7] focus on action discovery without any external guidance. For the application in hand, it is assumed that the user knows what the different actions are, leading to a choice of supervised techniques.
Works by [10, 5, 4, 6, 7] subsume large portions of the activity classification literature, and are an excellent source of reference.

3. DATASET

Videos from a static camera in a camera network are used for experiments. Each video is of duration equal to 20 minutes, giving a total footage time of 100 minutes. Using the dataset, 25 clips of pedestrians, 45 clips of skateboarders, and 45 clips of bikers were manually cropped. Each clip was of duration between 3-5 seconds. Since videos are variable frame rate, approximately 80 frames made up each video clip. All clips (total of 7000 frames) were manually annotated for the object of interest. In other words, annotating each clip gives a track. In order to automate the process, an ensemble tracker [1] using gentleboost was also employed.
The dataset used for this work presents several challenges. It simultaneously presents the following scenarios:

- Large scale variations: Scale of the object varies across its lifetime in the video. A 3-pixel man appears, changes gradually to a 30, 300 - pixel man and finally disappears from the field of view of the camera.
- Illumination Variation: Feeds are captured at varying times of the day, because of which there is severe illumination within and across videos.
- Occluding Objects: The dataset consists of lampposts that temporarily occlude the objects being tracked.
- Crowd Clutter: During late mornings/early afternoons, there is increased activity because of which people surrounding the target may contribute to severe clutter.
- Variable frame rate data: The frame rate of the videos is not constant. This is to be expected in surveillance videos.
- Variable quality data: Since data is being acquired wirelessly, there is a possibility of frame drops due to network traffic, and video quality does not always remain constant.

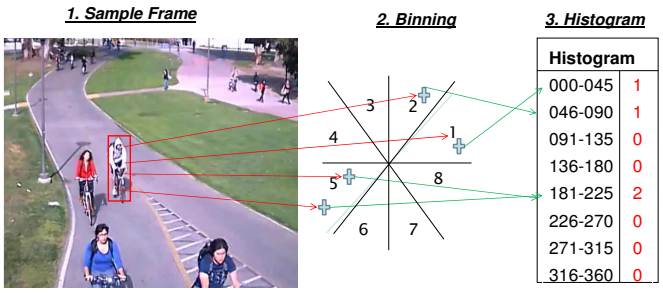

Figure 2: *Construction of an Oriented Optical Flow Histogram.*

Most importantly, the dataset presents a real world scenario without exercising any control over the environment on which the data is captured. Optical Flow histograms are not directly used for classification as it is a directional feature and will not generalize for classification, see Figure 1. Hence, features which are direction independent and having a significant physical meaning to them are computed.

4. LEARNING OPTICAL FLOW HISTOGRAM STATISTICS

The following section initially describes construction of the basic STOFS descriptor, after clarifying notations. Subsequently, techniques to make the descriptor scale-invariant and parts-based are discussed.

4.1 Notations and Nomenclature

In order to simplify the following discussion, we define important terms and clarify our notations.

Definition 1. 1. A track is the collection of an object's characterization throughout the length of a video clip. Subtracks are uniformly sampled subsets of tracks, and are of fixed size.

Definition 2. 1. A trajectory is the collection of centroids of an object's characterization throughout the length of a video clip. Subtrajectories are uniformly sampled subsets of trajectories, and are of fixed size.

Definition 3. 1. Histograms of Oriented Optical Flow (Figure 2) (HOOF, denoted henceforth as $h \in R^d$) is computed

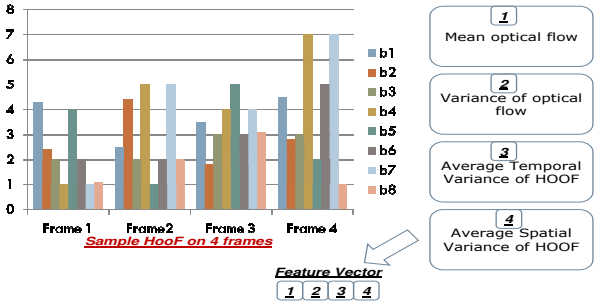

Figure 3: *Feature Extraction using HooF.*

over a patch, and is defined by the distribution of optical flow magnitudes over quantized orientations of flow vectors.

Each track is composed of several HOOFs and will be denoted as $Q_i(h) \in R^{d \times \sum_{j=1}^{N} T_{ij}}$, while the set of histograms constituting a sub-track will be denoted by $q_{ij}(h) \in R^{d \times T_{ij}}$. T_{ij} denotes the duration of the j^{th} sub-track extracted from the i^{th} video.

4.2 Statistics of Optical Flow Histograms

Since optical flow is a vector field, each flow vector has an angular orientation associated with it. The flow magnitudes are binned to their corresponding orientations to construct a histogram (shown as a table in figure 2). In this work, the number of bins is always fixed to 8. To ensure proper feature comparisons between bounding boxes of different sizes, all regions on which HOOF is extracted are initially normalized to a common size.

The following assumes a single sub-track to be denoted by $q = q_{ij}$ (subscripts are dropped to avoid notational clutter), where q is a matrix containing optical flow histograms along its columns. Without loss of generality, assume L to be the number of frames constituting a sub-track.

The features extracted from HOOF constructed from each sub-track are, see figure 3:

- Mean Optical Flow: The mean of all HOOF bins across a sub-track measures the velocity with which a target moves.

$$\mu = \frac{\sum_k \sum_l q(k,l)}{|q|}; 1 \leq k \leq d, 1 \leq l \leq L \quad (1)$$

- Average Variance of Optical Flow: Measures the smoothness of of flow of an object. Since this feature is capable of capturing if an object suddenly increases or decreases speed, it can be thought of as an acceleration cue.

$$\sigma = \frac{\sum_k \sum_l (q(k,l) - \mu)^2}{|q| - 1}; 1 \leq k \leq d, 1 \leq l \leq L \quad (2)$$

- Average Intra Frame Variance: Variance of optical flow vectors within a frame captures how the object deforms statically. Averaging this measure across all frames in a sub-track characterizes the amount of static deformation an object is capable of undergoing. It can be denoted as var_q.

$$\begin{aligned} q &= [h_1|h_2...|h_L] \\ var_q &= \frac{\sum_l var(h_l)}{L}; 1 \leq l \leq L \end{aligned} \quad (3)$$

- Average bin-wise Variance: Average variance of optical flow vectors of each bin across a sub-track measures temporal deformation of an object. In other words, it is a measure of the ability of an object to deform along every orientation across time. It can be denoted as $var_{(q^T)}$.

$$\begin{aligned} q^T &= [g_1|g_2...|g_d] \\ var_{q^T} &= \frac{\sum_k var(g_k)}{d}; 1 \leq k \leq d \end{aligned} \quad (4)$$

The resulting 4-D feature vector $[\mu_{ij}, \sigma_{ij}, var_{q_{ij}}, var_{q_{ij}^T}]$, thus captures a notion of spatio-temporal deformation (*deformotion*) of an object.

4.3 Depth Normalization

In the dataset under consideration, there is a huge disparity between bicyclists moving close to the camera, and the skateboarders/pedestrians moving far away from the camera. It would be extremely useful to learn an approximate depth map which could normalize the scale variant features extracted from the video. We propose to make use of the output of the tracker (perimeter of bounding boxes) to achieve fast and approximate depth normalization. The basic idea is that an object's perimeter is low at a large distance from the camera, while the same object's perimeter is higher when it is closer to the camera. Using the object perimeters as data points, a Nadaraya-Watson kernel regression using gaussian kernels (K_h) is performed to obtain an approximate velocity/depth normalization map, see Figure 4. This map is utilized for normalization of features extracted from the video clips. In Equation 5, x is the set of pixel coordinates in the image, X_i is the set of n points on which bounding boxes are available, Y_i is the perimeter of the bounding boxes. Note that $\hat{m_h}(x)$ is the approximate depth map used for normalizing scale variant features.

$$\hat{m_h}(x) = \sum_{i=1}^{n} \frac{K_h(x - X_i)}{\sum_{j=1}^{n} K_h(x - X_j)} Y_i \quad (5)$$

4.4 Parts-Based STOFS

Research in visual psychology advocates a parts based nature of human perception [8], following which many object recognition methods employ parts-based models. This work introduces parts based STOFS (p-STOFS), similar to [8], the main difference is the manner in which parts are constructed. [8] attempts to recognize human actions by breaking the detected human silhouette into semantically meaningful parts and learning discriminative features on them. In contrast, the proposed method constructs parts for any action of interest by searching for parts that maximize discriminative ability across classes, while parts are not constrained to be semantically meaningful. In the present work, the part being searched for is a line $l' \in \{0, H\}$ that splits the rectangular bounding box into two blocks (P_1, P_2). H refers to the

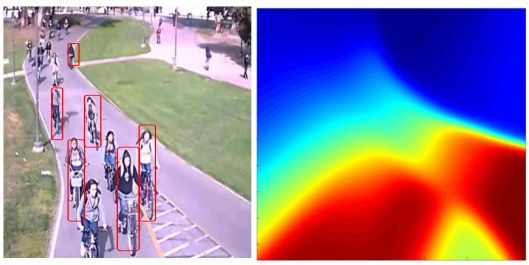

Figure 4: *[Best Viewed in Color] Depth/Velocity Normalization, (left) a frame from the video having severe depth disparity. (right) normalized velocity map using kernel regression*

Table 1: Confusion Matrix of Discriminant Analysis

Conf.Mat	Bike	Skate	Person
Bike	.53	.46	0
Skate	.2	.66	.13
Person	0	0	1

height of a bounding box under consideration. In principle it is possible to recursively split bounding boxes by minimizing a discriminative cost, however for experiments only a single split is employed.

5. EXPERIMENTS

Experiments are conducted on the dataset described in Section 3. A naive trajectory feature employing centroids of bounding boxes is employed as a baseline method. Various features of centroid motion were tested on the baseline (velocity, acceleration, Fourier and Wavelet Spectrums, turning angles). It was found that accelaration provided consistent results and was hence employed for baseline comparisons. Subsequently, STOFS and p-STOFS are employed for classification on the same data. Finally, a comparison to Sparse Spatio Temporal Features (SSTF), a state of the art action descriptor is presented.

5.1 Sub-Track Classification using STOFS

Experiments were performed separately on bounding boxes obtained from manual annotations (410 in total, split into 160 training and 250 test tracks) and the ensemble tracker (233 in total, split into 160 training and 73 test tracks). The naive method was initially employed for the three-way classification. Confusion matrix resulting from employing the baseline is shown in Table 1. As can be observed, sub-tracks of pedestrians are easily separated from bikers and skateboarders without any confusion. Hence one can safely assume the separability of pedestrians using optical flow magnitudes. However, there is severe confusion between bikers and skateboarders, see Table 1 and 2.

Table 2: Baseline Accuracy

Method	Mean Accuracy(S/B)	Mean Accuracy(S/B/P)
DA	61.05 ± 7.81	74.38 ± 3.8
SVM	68.6 ± 7.93	79.66 ± 2.81

Table 3: STOFS Accuracy - SubTrack Classification

Method	Accuracy(Tracker)	Accuracy(Manual)
LDA	84.98 ± 1.39	83.95 ± 1.39
QDA	85.37 ± 3.41	83.81 ± 1.57
SVM	86.56 ± 1.26	83.41 ± 1.64
Boosting	86.30 ± 3.64	81.99 ± 2.19

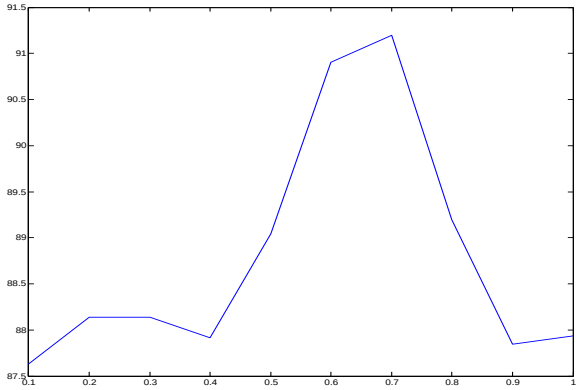

Figure 5: *[Best Viewed in Color] Classification errors as the line splitting the bounding box into parts is moved. X-axis corresponds to the position of the line l' splitting the bounding box into parts, Y-axis corresponds to the classification accuracy for varying values of l'.*

STOFS is extracted as a "bag of features" over which classifiers are trained during the training phase. Every new STOFS descriptor arriving at test time is projected onto the trained classifier for a class decision. The accuracy of sub-track classification using STOFS is shown in Table 3. It must be noted that different classifiers were employed only for completeness of experiments, and the focus is on how well STOFS can be handled by these different methods.

5.2 Track Classification from Sub-Track labels

Since the ultimate goal is to label tracks, a principled method to infer track labels from sub-track labels is required. Though more complex time series inference (HMM) could be employed, we propose a simple and efficient polling procedure, based on most frequent sub-track label in every track. Experiments on track label inference using the proposed simple technique yield considerable improvement in overall classification accuracy. Manually annotated bounding boxes were classified with an accuracy of 95.4% using boosting (Table 4), while bounding boxes returned by the ensemble tracker were classified with an accuracy of 96.6% by a boosted classifier (Table 4).

Table 4: STOFS Accuracy - Track Classification

Method	Accuracy(Tracker)	Accuracy(Manual)
LDA	90.00	90.80
QDA	96.60	90.90
SVM	96.60	91.95
Boosting	96.60	95.4

Table 5: Confusion Matrix of Sparse-Spatio Temporal Features

Conf.Mat	Bike	PersonSkate
Bike	1	0
PersonSkate	.6	.4

Table 6: STOFS vs p-STOFS - Sub-Track Classification (ONLY BIKE VS SKATEBOARDERS)

Method	STOFS Accuracy	p-STOFS Accuracy
LDA	77.75 ± 1.92	79.12 ± 1.85
QDA	77.35 ± 1.55	79.70 ± 1.94
SVM	75.59 ± 2.15	79.44 ± 2.24
Boosting	77.46 ± 1.56	77.13 ± 1.91

5.3 Comparison to SSTF

Since the proposed method is a feature descriptor for activity classification, it is tested against a state of the art activity descriptor (SSTF) [4], using default parameters of their implementation. Since SSTF is an unsupervised method for activity classification, the data is modified for comparative experiments (videos cropped to contain single activity essentially reducing search space as SSTF does not use trackers). The bike paths in the video are cropped from the sidewalk, and SSTF is trained to distinguish "bikepath" videos from "sidewalk" videos. The test is to ascertain whether SSTF could classify a biking activity from a (walking / skateboarding) activity. As the confusion matrices indicate, SSTF does not provide a very good distinction, see Table 5. It must be noted that SSTF is an established method proven to work well in activity classification. Hence, the comparison justifies promise of the proposed method. Tests on more generic datasets considering parameter variations of competing methods is part of future work.

5.4 Parts-Based STOFS

It was also observed in experiments that parts-based STOFS fared better than the STOFS feature alone. Since classifying bikers from skateboarders is the tougher problem, we concentrate on this 2-way classification while comparing STOFS with p-STOFS. For the training phase, 101 subtracks each of bikers and skateboarders were used for training, while 137 subtracks each of bikers and skateboarders were used for testing. Observe a gain in performance while employing p-STOFS, in comparison to STOFS stand alone. See Tables 6 and 7.

5.5 Discussion

In both the testing phases the standard deviation is relatively low for all classifiers, highlighting the robustness of the features. Also, the single most important feature for classification between bikers and skateboarders is the average bin-wise variance of HOOF. The motion of skateboarders is non-smooth temporally as there are minor direction fluctuations. Moreover the skateboarders exercise the greater degree of freedom of motion as they do not have a fixed path to adhere to and need to avoid obstacles. Hence, this feature is on an average higher for skateboarders compared to bikers. The bikers have a slightly higher average intra-frame variance of HooF as they undergo more local deformations as their legs move while biking. But, this is not as significant as the temporal bin-wise variance feature. Factoring out the complexity of the optical flow and tracker computations, the testing complexity of the proposed algorithm is linear $O(N)$ in the number of training samples (N). The algorithm takes about a second on a 640 by 480 frame (optical flow, feature construction, and projection onto trained classifiers) in Matlab.

Table 7: STOFS vs p-STOFS - Track Classification (ONLY BIKE VS SKATEBOARDERS)

Method	STOFS Accuracy	p-STOFS Accuracy
LDA	86.30 ± 3.45	89.10 ± 3.21
QDA	86.64 ± 4.01	88.88 ± 3.48
SVM	86.05 ± 3.81	89.85 ± 3.11
Boosting	81.86 ± 3.75	84.83 ± 3.60

6. CONCLUSIONS

In summary, a methodology for action classification using motion cues alone is proposed. Learning appearance cues could yield a significant boost to classification accuracy, but could prove difficult to generalize. The striking aspect of this work is action recognition in a real world setting with a very low dimensional "motion" descriptor, without reliance on appearance features. Future work includes graphical model based track inference, and testing the viability of pyramid match kernels for optical flow histograms.

7. ACKNOWLEDGMENTS

The authors gratefully acknowledge support from NSF III-0808772, ONR DURIP-N00014-08-1-0791 and NIMH 1 R01 MH070539-01.

8. REFERENCES

[1] S. Avidan. Ensemble tracking. *PAMI*, pages 261–271, 2007.

[2] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. *CVPR*, 2009.

[3] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. *ECCV 2006*, pages 428–441, 2006.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *ICCV VS-PETS 2005*, pages 65–72, 2005.

[5] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *ICCV*, 2003.

[6] I. Laptev. On space-time interest points. *IJCV*, 64(2):107–123, 2005.

[7] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318, 2008.

[8] D. Ramanan, D. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *PAMI*, pages 65–81, 2007.

[9] S. Roth and M. Black. On the spatial statistics of optical flow. *IJCV*, 74(1):33–50, 2007.

[10] L. Zelnik-Manor and M. Irani. Statistical analysis of dynamic actions. *PAMI*, pages 1530–1535, 2006.