Optical Acceleration for Motion Description in Videos

Anitha Edison & Jiji C. V.
College of Engineering, Trivandrum
Kerala, India
anithaedison@cet.ac.in, jijicv@cet.ac.in

Abstract

Modern techniques for describing motion in videos are centred around velocity descriptors based on optical flow. Realizing that acceleration is as important as velocity for describing motion information, in this paper first we propose two different algorithms to compute optical acceleration. Delving deeper into the concept of optical acceleration, we use two descriptors: histogram of optical acceleration (HOA) and histogram of spatial gradient of acceleration (HSGA), to effectively encode the motion information. To assess the effectiveness of these descriptors for motion encoding, we applied it for human action recognition and abnormal event detection in videos. In fact, we used acceleration descriptors in conjunction with velocity descriptors to get a better description of motion in videos. Our experiments reveal that acceleration descriptors could provide additional information that velocity descriptors missed and hence combining them results in a superior motion descriptor.

1. Introduction

Advent of cheaper video capture and storage devices resulted in large volume of video data, making automated video analysis a necessity. Estimating motion between frames is the key to video analysis.

Researchers were always intrigued by the way human brain perceives motion. It was this fascination that lead to the introduction of the concept of optical flow by James J. Gibson, a psychologist. He analysed the use of optical flow by human brain to recognise motion[9]. Computer vision researchers in an attempt to mimic the working of a human brain, introduced a motion vector called optical flow; to represent the velocity distribution of moving objects in an image sequence[10]. Researchers across the globe are still trying to improve motion estimation using optical flow, among which computation of optical flow using semantic segmentation [23], convolutional neural networks [6] etc. are some recent examples.

Techniques involving optical flow are used in diverse areas of computer vision. Motion estimation using optical flow is used to detect moving objects in video [3]. Motion analysis techniques are used to remove temporal redundancy in videos [24] for video coding. Automated video surveillance uses motion information obtained from optical flow for anomaly detection [29], video summarization [16] and action recognition [25]. Motion vectors obtained using optical flow finds application in medical imaging, to estimate blood flow from MRI image sequences [20]. These researches depict the importance of optical flow and motion estimation in computer vision. Any attempt to improve motion estimation by optical flow will benefit the aforementioned areas significantly.

In this paper we present two new methods for computing ‘optical acceleration’ to supplement the motion information provided by optical flow. We derive two motion descriptors, histogram of acceleration (HOA) and histogram of spatial gradient of acceleration (HSGA), from it for video analysis. We propose an alternative approach that uses acceleration descriptors along with velocity descriptors to give a better description of motion in videos. We evaluate this approach in the context of action recognition and abnormal event detection. Our action recognition system uses HSGA in conjunction with velocity descriptor - motion boundary histogram (MBH), to train a classifier that can recognize actions performed in challenging environments. We have used a concatenated feature set consisting HOF, MBH, HOA and HSGA descriptors in a sparse reconstruction framework for anomaly detection. Experiments performed on standard datasets show that the descriptor combination significantly improves the recognition rate when compared to using either of them alone.

Rest of the paper is organised as follows. Section 2 reviews some common motion descriptors. In section 3 we derive two different algorithms for the computation of optical acceleration. Section 4 details the descriptors derived from optical acceleration. Use of optical acceleration for action recognition is described in section 5 and for anomaly detection in section 6. Section 8 concludes the paper.

2. Motion Descriptors

This section reviews some of the prominent feature descriptors used for describing motion in videos for various computer vision applications.

Success of histogram of oriented gradients (HOG) for object detection in static images, urged Klaser *et al.* to build on it and extend the descriptor to time domain to form histogram of oriented 3D spatio-temporal gradients (HOG3D) [12]. SURF and SIFT descriptors where also extended to time domain to obtain extended speeded up robust feature (Extended SURF) [26] and 3-Dimensional SIFT respectively. A spatio-temporal gradient was used for anomaly detection in extremely crowded scenes by Kratz *et al.* [13]. This kind of extensions of spatial features to time domain have the disadvantage of treating the space and time domains equally and therefore is not very apt representation.

Laptev *et al.* used a combination of histogram of optical flow (HOF) and histogram of oriented gradients (HOG) to describe scene and motion respectively, for action recognition [14]. Dalal *et al.* introduced motion boundary histogram (MBH) which encodes the velocity in object boundaries for human detection in videos, and it was adopted by Wang *et al.* [25] for action recognition. Dense trajectories obtained by tracking sample points taken on a regular grid were used for video analysis [25]. A multi-scale histogram of optical flow (MHOF) which consist of histogram of optical flow quantised to 16 bins which include two scales of 8 bins each was successfully used for abnormal event detection[4]. Histogram of maximal optical flow projections (HMOFP) is another variation of HOF which was used for anomaly detection in crowded scene [15]. HMOFP as the name suggest is obtained by projecting the optical flow vectors in each bin to angle bisector of that bin and choosing the magnitude of maximal projection as descriptor corresponding to each bin. All these descriptors are derived from optical flow and encode velocity information.

3. Optical Acceleration

When an object moves with varying velocity there will be change in optical flow. Optical acceleration is defined as the rate of change of optical flow and it gives apparent acceleration of each pixel in a frame. We speculate that optical acceleration helps in motion description based on the fact that physical acceleration can differentiate motion as uniform, accelerated and decelerated motion.

Acceleration component of motion was first used for anomaly detection by Nallaivarothayan *et al.*[19]. They used it effectively to detect non pedestrian entities in walkways, but this may not be the case with many other scenarios like action recognition. The reason being they neglect the direction of velocity change and so their descriptor is not fully suitable to represent motion. Another interesting work in this regard used a second order differential of original image, called acceleration stream, as an input to multi stream CNN for action recognition[11].

We develop two approaches for computation of optical acceleration and derive descriptors from it giving due consideration to the direction of change in velocity. We have named them Horn- Schunck optical acceleration and Farneback optical acceleration, after optical flows on which we based the derivations of acceleration.

3.1. Horn-Schunck Optical Acceleration

It is derived by applying brightness and smoothness constraint similar to the case of optical flow[10].

For an image with brightness $E(x, y)$ at point (x, y), the optical flow constraint equation (equation 1) can be differentiated in time to obtain what we call the 'optical acceleration constraint'(equation 2). (Abbreviations E_x, E_y and E_t stands for partial derivatives of E(x,y) with respect to x, y and t respectively, (u, v) is optical flow vector and (a, g) the acceleration vector)

$$uE_x + vE_y + E_t = 0 \tag{1}$$

$$aE_x + gE_y + [u\frac{\partial E_x}{\partial t} + v\frac{\partial E_y}{\partial t} + \frac{\partial E_t}{\partial t}] = 0 \tag{2}$$

An additional constraint that the neighbouring pixels will have same rate of change of velocity gives the smoothness constraint:

$$\min\ a_x^2 + a_y^2\ and\ \min\ g_x^2 + g_y^2 \tag{3}$$

a_x, a_y, g_x, g_y are the derivatives of a and g with respect to x and y. The problem of computing optical acceleration becomes the one of minimising error in optical acceleration constraint equation (ξ_{ac}) subjected to smoothness constraint equation (ξ_{sc}). Error minimization problem is formulated as

$$\min\ \xi^2 = \iint (\xi_{sc}^2 + \lambda^2 \xi_{ac}^2) dxdy \tag{4}$$

where

$$\xi_{ac}^2 = [uE_x + vE_y + E_t]^2 \tag{5}$$

and

$$\xi_{sc}^2 = a_x^2 + a_y^2 + g_x^2 + g_y^2 \tag{6}$$

Solution for the minimisation problem in 4 is obtained as in equation 7 and 8 using calculus of variation, where $E_{xt} = \dfrac{\partial E_x}{\partial t}$, $E_{yt} = \dfrac{\partial E_y}{\partial t}$ and $E_{tt} = \dfrac{\partial E_t}{\partial t}$

$$\bigtriangledown^2 a = \lambda[aE_x^2 + gE_xE_y + uE_xE_{xt} + vE_xE_{yt} + E_xE_{tt} \tag{7}$$

$$\bigtriangledown^2 g = \lambda[aE_xE_y + gE_y^2 + uE_yE_{xt} + vE_yE_{yt} + E_yE_{tt} \tag{8}$$

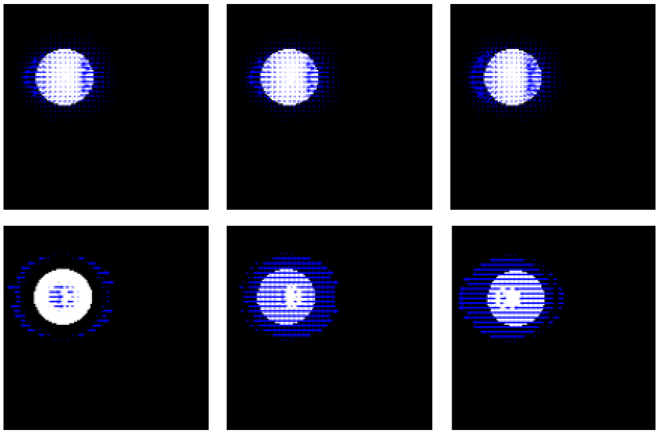

Figure 1. Optical acceleration for horizontal movement of a circular disc in Left: uniform motion, Middle: acceleration, Right: deceleration Top: using Horn Schunck optical acceleration bottom: using Farneback optical acceleration.

Approximating the Laplacian of a and g for discrete domain gives an iterative solution for acceleration shown in equation 9 and 10.

$$a_{i,j}^{(n+1)} = \overline{a}_{i,j}^{(n)} - \frac{\lambda E^{(n)} E_x}{1 + \lambda(E_x^2 + E_y^2)} \tag{9}$$

$$g_{i,j}^{(n+1)} = \overline{g}_{i,j}^{(n)} - \frac{\lambda E^{(n)} E_y}{1 + \lambda(E_x^2 + E_y^2)} \tag{10}$$

where

$$E^{(n)} = [a_{i,j}^{(n)} E_x + g_{i,j}^{(n)} E_y + u_{i,j} E_{xt} + v_{i,j} E_{yt} + E_{tt}] \tag{11}$$

and

$$\overline{a}_{i,j}^{(n)} = \frac{1}{4}[a_{i-1,j}^{(n)} + a_{i+1,j}^{(n)} + a_{i,j-1}^{(n)} + a_{i,j+1}^{(n)}] \tag{12}$$

Acceleration computed using this method has the inherent disadvantages of global optical flow methods, since acceleration will always be local and the assumptions used here are global in nature.

3.2. Farneback Optical Acceleration

Farneback optical acceleration is estimated by expanding each frame as a quadratic polynomial. Any local signal can be expressed as a quadratic polynomial in equation 13[8].

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c \tag{13}$$

where **x** is a local coordinate, **A**: a symmetric matrix, **b**: a vector and c: a scalar. Image intensity around a $N \times N$ neighbourhood of a point $\mathbf{x} = [x, y]^T$ in a frame can be expanded using polynomial basis set $\mathbf{B} = \{1, x, y, x^2, y^2, xy\}$ as

$$E(x, y) = C_1 + C_2 x + C_3 y + C_4 x^2 + C_5 y^2 + C_6 xy \tag{14}$$

Substituting $\mathbf{x} = [x, y]^T$in equation 13 and comparing it with 14 gives

$$c = C_1,\ \mathbf{b} = \begin{pmatrix} C_2 \\ C_3 \end{pmatrix},\ \mathbf{A} = \begin{pmatrix} C_4 & \frac{C_6}{2} \\ \frac{C_6}{2} & C_5 \end{pmatrix} \tag{15}$$

The set of coefficients, $\mathbf{C} = [C_1, C_2, C_3, C_4, C_5, C_6]$, can be estimated by normalized convolution with the polynomial basis **B** (equation 16). In the equation, **f(x)** is image intensity, * represents conjugate transpose and $\mathbf{W_a}$ and $\mathbf{W_c}$ are diagonal matrices corresponding to certainty of signal and applicability of basis respectively.

$$\mathbf{C} = (\mathbf{B}^* \mathbf{W_a} \mathbf{W_c} \mathbf{B})^{-1} \mathbf{B}^* \mathbf{W_a} \mathbf{W_c} \mathbf{f(x)} \tag{16}$$

If image intensity at neighbourhood of a pixel in two consecutive frames in a video is approximated using a set of coefficients $\mathbf{A}_1, \mathbf{b}_1, c_1$ and $\mathbf{A}_2, \mathbf{b}_2, c_2$, then the displacement(**d**) of the pixel can be obtained using equation 17 [8]

$$\mathbf{d} = -\frac{1}{2} \mathbf{A}_1^{-1} (\mathbf{b}_2 - \mathbf{b}_1) \tag{17}$$

Difference between two such displacements between three adjacent frames gives optical acceleration (a, g).

To evaluate the two optical acceleration methods we have conducted experiments on a synthetic video of a circular disc moving horizontally to the right. For Horn Schunck method the parameters λ and n were chosen as .01 and 100 respectively and a 3×3 neighbourhood was chosen for Farneback optical acceleration. Figure 1 shows the optical acceleration between three consecutive frames. Top row shows the Horn Schunck optical acceleration while the bottom is the Farneback optical acceleration. Left, right and middle are optical acceleration corresponding to uniform motion, deceleration and acceleration respectively. It is evident from the figure that the second method can differentiate uniform, accelerated and decelerated motion quite well.

Visual evaluation proved Farneback method to be more effective than Horn Schunck for representing acceleration, so we have used Farneback acceleration to extract acceleration descriptor for encoding motion information.

4. Acceleration Descriptors

A descriptor based on optical acceleration was first used by Nallaivarothayan *et al.* [19] to detect speed related anomalies in surveillance videos. They have incorporated acceleration information by taking time derivative of magnitude image of optical flow for each frame. Their acceleration descriptor neglects the direction of change in velocity and hence is not perfectly suited for motion description. In this section, we detail two descriptors histogram of optical acceleration (HOA) and histogram of spatial gradient of acceleration (HSGA) derived from optical acceleration. As in the case of high speed anomalies, these descriptors which encode acceleration can be used effectively to differentiate motion with varying acceleration.

4.1. Histogram of Optical Acceleration

Histogram of Optical Acceleration (HOA) is derived from Farneback optical acceleration. If the horizontal and vertical components of acceleration of a point (x, y) in frame t is given by $a_{(x,y)}^{(t)}$ and $g_{(x,y)}^{(t)}$ respectively, the magnitude and orientation of acceleration of the point given by equation 18 and 19 is calculated. This is computed for each point in a frame to form acceleration magnitude and orientation images.

$$\theta_{OA} = tan^{-1}\left[\frac{g_{(x,y)}^{(t)}}{a_{(x,y)}^{(t)}}\right] \tag{18}$$

$$|OA| = \sqrt{\left|a_{(x,y)}^{(t)}\right|^2 + \left|g_{(x,y)}^{(t)}\right|^2} \tag{19}$$

Interest region in the video is divided into blocks of size $M \times M \times K$ which is further divided into $n_x \times n_y \times n_t$ cells. For each cell, orientation of optical acceleration (θ_{OA}) is quantized into 8 bins and a histogram weighted by its magnitude ($|OA|$) is computed. Values are linearly interpolated between two neighbouring bins using bilinear interpolation. Histogram of cells in a particular block are concatenated and normalised to form HOA. HOA has the advantage of reduced descriptor size and this makes clustering easier for bag of words encoding and reduces storage space.

4.2. Histogram of spatial gradient of acceleration

The descriptor HOA encodes the acceleration of the whole frame, but the acceleration along object boundaries is more important than acceleration component as such. Descriptor histogram of spatial gradient of acceleration (HSGA) is introduced to encode the acceleration along the boundaries[7]. HSGA uses spatial derivative of horizontal and vertical components of acceleration to capture acceleration changes in the boundaries. For a point (x, y) in frame t, spatial derivatives of horizontal component of acceleration are computed using Sobel operator as in equation 20 and 21, where $.*$ stands for element by element multiplication.

$$A_x = \begin{bmatrix} a_{x-1,y-1}^t & a_{x,y-1}^t & a_{x+1,y-1}^t \\ a_{x-1,y}^t & a_{x,y}^t & a_{x+1,y}^t \\ a_{x-1,y+1}^t & a_{x,y+1}^t & a_{x+1,y+1}^t \end{bmatrix} .* \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{20}$$

$$A_y = \begin{bmatrix} a_{x-1,y-1}^t & a_{x,y-1}^t & a_{x+1,y-1}^t \\ a_{x-1,y}^t & a_{x,y}^t & a_{x+1,y}^t \\ a_{x-1,y+1}^t & a_{x,y+1}^t & a_{x+1,y+1}^t \end{bmatrix} .* \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{21}$$

Magnitude and orientation of spatial gradient of acceleration for the point (x, y) is given by equation 22 and 23 and this is computed for all points in a frame.

$$\theta_A = tan^{-1}\left[\frac{A_y}{A_x}\right] \tag{22}$$

$$|A| = \sqrt{|A_x|^2 + |A_y|^2} \tag{23}$$

As in case of HOA , the region around feature point, is divide into space time volume of size $M \times M \times K$, which is further divided into $n_x \times n_y \times n_t$ cells. The orientation of the gradient (θ_A) in a cell is quantized into histogram of 8 bins weighted by magnitude ($|A|$). Histogram of cells in a block are concatenated and normalized to form histogram of spatial gradient of horizontal component of acceleration. The descriptor is computed in a similar manner for vertical component of acceleration.

5. Action recognition

An important application of motion estimation is in action recognition and it has drawn a lot of attention in recent years. This can be attributed to its variety of applications in automated video surveillance, human computer interface, content based video indexing etc. Past decade saw researchers use a wide variety of techniques ranging from statistical models like Hidden Markov Models (HMM)[27] to deep learning networks like Convolutional Neural Networks (CNN)[28] for action recognition. Local feature based methods[12] and template matching methods[2] which were used for object recognition, hand writing recognition etc. were also extended for action recognition. Of all action recognition techniques, local feature based methods are the most popular. An action recognition system using local features consists of extracting features from videos to train a classifier, which is then used

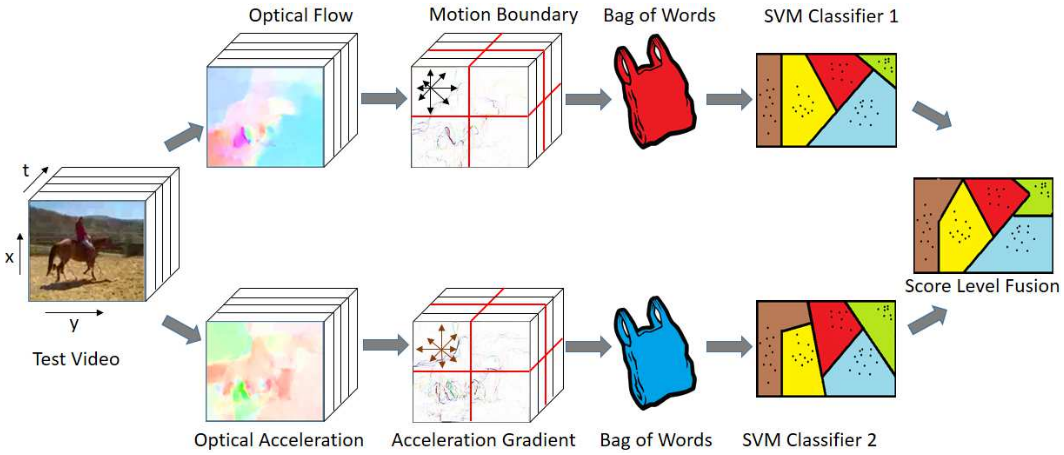

Figure 2. Flow diagram of proposed action recognition system

to recognize action in a given video. Among existing motion descriptors for action recognition, motion boundary histogram (MBH) computed around improved dense trajectories (IDT) [25] performs the best. In this section we showcase the use of acceleration descriptor HSGA to supplement the performance of MBH for action recognition.

5.1. Optical acceleration for action recognition

Figure 2 shows the flow diagram of our proposed approach where we fuse optical acceleration descriptor, HSGA [7] with optical velocity descriptor MBH [25] for action recognition. Optical flow and optical acceleration are computed for the entire video. Spatial gradient of horizontal and vertical components of acceleration and flow are used to capture the acceleration and velocity information along the boundaries. HSGA and MBH for each video is calculated and they are encoded using bag of features encoding. Encoded bag of features corresponding to the two descriptors are used to train separate SVM classifiers. Most commonly used descriptor fusion method is representation level fusion in which different descriptors encoded using bag of features are concatenated to form fused descriptor. But when the descriptors have totally different characteristics concatenating them will not be much effective [7]. So we are introducing SVM score level fusion which will be more effective in case of complementing descriptors. SVM score level fusion consists of training two separate SVM classifiers, one for MBH and other for HSGA, and fusing their probability score to obtain proper recognition. Consider training two different SVMs S_h and S_m using HSGA and MBH respectively, and let the corresponding test probability scores for n action categories be $P_h = \{p_h^1, p_h^2, ..., p_h^n\}$ and $P_m = \{p_m^1, p_m^2, ..., p_m^n\}$. According to the way, in which the score are fused SVM score level fusion can be classified into *max fusion*, *mean fusion* and *hybrid max-mean fusion*. In max fusion, $P = max(P_h, P_m)$ is obtained and action corresponding to $max(P)$ is chosen as correct action class. $P = mean(P_h, P_m)$ is obtained in mean fusion and n corresponding to $max(P)$ is the recognised action class. Max fusion and mean fusion methods are combined in hybrid max-mean fusion method. In hybrid fusion, $P_T = \{p_T^1, p_T^2, ..., p_T^n\} = max(P_h, P_m)$ is calculated and equation24 is used to find P. As in case of previous methods, action corresponding to maximum P is chosen.

$$P = \begin{cases} max(P_h, P_m), & p_T^i > T \,\forall\, i = 1, 2, ..., n \\ mean(P_h, P_m), & otherwise \end{cases} \tag{24}$$

where T is an experimentally chosen threshold.

5.2. Experiments and Discussions

Three standard action recognition datasets namely, KTH, YouTube and UCF50 action recognition datasets were used to evaluate the performance of our fused descriptor. KTH dataset [22] consists of a total of 599 videos of six human actions shot in lab. 50 videos from each action class were used for training and remaining were used for testing. YouTube [17] and UCF50 [21] datasets are more challenging datasets compared to KTH, since the videos have large variations in scale and view point, background motion, and camera motion, etc. YouTube action dataset contains 1168 videos divided into 11 action categories while UCF50 is an extension of YouTube dataset with 6666 videos divided into 50 action categories. Videos in each action class are grouped into 25 groups and the evaluation set-up used is ‘leave one out cross validation’. Figure 3 shows some sample frames from the datasets.

To evaluate the performance of our motion descriptors, we use the improved dense trajectories framework adopting parameters from [25]. Motion descriptors were computed around trajectories for a spatio-temporal volume of size 32 × 32 × 15 divided into 2 × 2 × 3 cells. Farneback acceleration with a neighbourhood size of 7 × 7 was used to

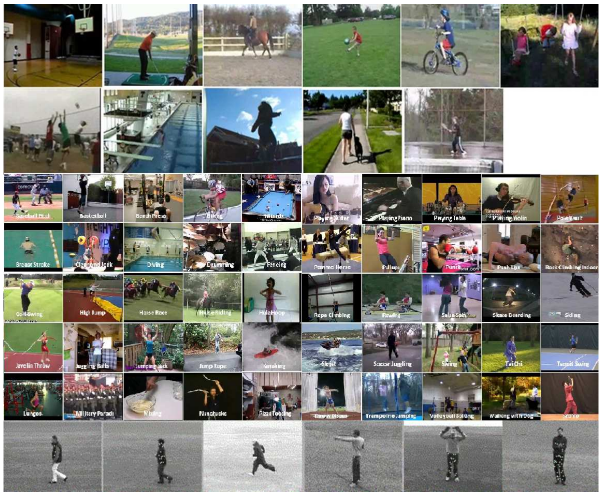

Figure 3. Sample frames from datasets

extract acceleration descriptors. Bag of features approach with a dictionary size of 4000 was used for encoding and support vector machine with χ squared kernel was used for classification.

Dataset	HOA	HSGA
KTH	90.97 %	95.32 %
YouTube	72.38 %	76.51%
UCF50	62.23%	69.97 %

Table 1. Performance of acceleration descriptors

We first evaluate the performance of our acceleration descriptor on the three datasets. It is evident from Table 1 that HSGA performs better than HOA on all datasets. The results lead to the conclusion that acceleration information along the object boundaries is more relevant for action recognition, than the acceleration information as such. So we propose to combine our acceleration descriptor HSGA with its velocity counterpart MBH.

Dataset	Representation	Mean	Max	Hybrid
YouTube	82.68%	83.86%	83.46%	**84.06%**
UCF50	76.70%	**78.25%**	77.6%	78.21%

Table 2. Comparison of different fusion Schemes

We have experimented on different methods for fusing HSGA with MBH and the results of fusion using representation level fusion and SVM score level fusion methods are presented in Table 2. SVM fusion methods give an increased recognition accuracy, among which hybrid fusion method performs the best with a recognition accuracy of 84.06% on YouTube dataset and 78.21% on UCF50 dataset. Experiments were conducted to study the impact of threshold on recognition accuracy of hybrid fusion method and threshold was fixed at 0.55.

Dataset	HSGA	MBH	Fused
KTH	94.65 %	95.32 %	**96.32 %**
YouTube	76.51 %	82.88 %	**84.06 %**
UCF50	69.97 %	77.33 %	**78.21 %**

Table 3. Comparison of individual & fused descriptors

Table 3 shows the comparison of recognition accuracy of individual descriptors, HSGA and MBH, with the descriptor fused using hybrid method. Since HSGA and MBH have totally different properties, HSGA could recognize some actions that MBH could not and vice versa. Hence fusing the descriptors improved the recognition accuracy of the action recognition system. Figure 4 shows the confusion tables of YouTube dataset for individual descriptors and fused descriptor. Fusion of descriptors improved recognition rate of basketball shooting, diving, golf swinging, soccer juggling, swinging, tennis swinging and volleyball spiking compared to using either of the descriptors alone. Actions from UCF-50 dataset with increased recognition accuracy due to fusion is presented in table 4. For all other actions in the dataset, recognition accuracy of the fused descriptor is either same

Action	HSGA	MBH	Fused
Baseball Pitch	76.67%	75.33%	79.33%
Basketball	61.31%	60.58%	64.23%
Bench Press	85.63%	89.38%	91.88%
Golf Swing	82.73%	84.17%	87.77%
High Jump	71.54%	77.24%	78.05%
Hula Hoop	71.2%	69.6%	77.6%
Javelin Throw	44.44%	51.28%	52.14%
Jumping Jack	91.06%	90.24%	91.87%
Jump Rope	79.86%	85.42%	86.81%
Kayaking	73.25%	82.17%	84.08%
Military Parade	75.59%	78.74%	83.46%
Nunchucks	31.82%	50%	53.03%
Pizza Tossing	46.49%	55.26%	58.77%
Playing Guitar	68.13%	60.63%	76.25%
Playing Piano	70.48%	80%	82.86%
Playing Violin	71%	86%	87%
Pommel Horse	88.62%	93.5%	94.31%
Rock Climbing	72.22%	64.58%	75%
Rope Climbing	48.46%	58.46%	59.23%
Skate Boarding	70%	75.83%	76.67%
Skiing	58.33%	64.58%	65.97%
Skijet	61%	71%	75%
Swing	78.1%	88.32%	89.05%
Walking	56.1%	65.85%	67.48%
YoYo	64.57%	80.31%	81.1%

Table 4. Comparison of recognition accuracy for UCF 50 dataset

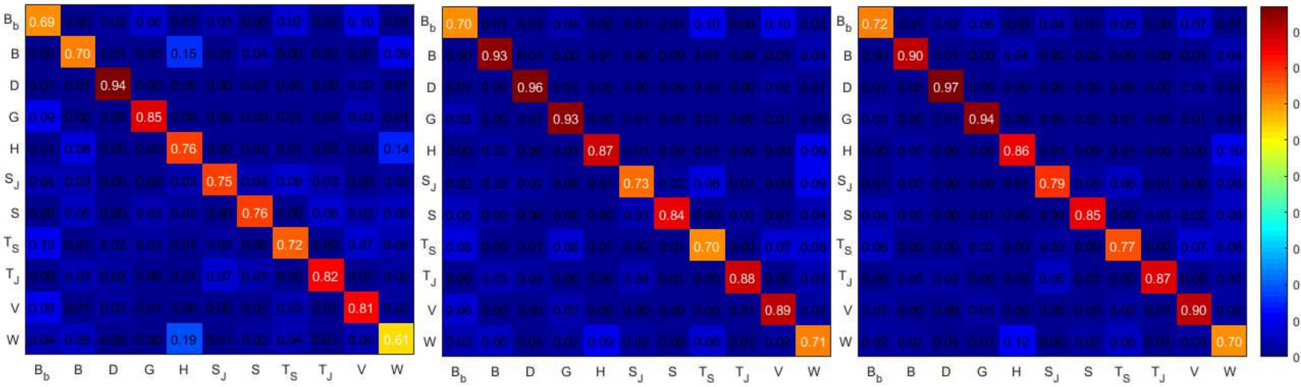

Figure 4. Comparison of recognition rate of individual and fused descriptors on YouTube dataset: confusion matrix for HSGA (left), MBH (middle) and fused MBH/HSGA(right). Labels denote: B_b: basketball shooting, B: biking/cycling, D: diving, G: golf swinging, H: horse back riding, S_J: soccer juggling, S: swinging, T_S: tennis swinging, T_J: trampoline jumping, V: volleyball spiking, and W: walking

or comparable with that of MBH. HSGA performs better than MBH for actions like baseball pitch, basketball, hula hoop, jumping jack, playing guitar, rock climbing etc and hence fusion of velocity and acceleration descriptors improves the recognition rate. Both figure 4 and table 4 show that there is a considerable increase in recognition accuracy for actions with significant acceleration information like golf swing, tennis swing and soccer juggling etc. From the above discussion it is evident that acceleration is as important as velocity to represent motion in videos and fusing acceleration and velocity descriptors compliment each other and improve representation further. Strength of HSGA lies in the fact that it could recognize some actions that MBH missed. A particular action is not always performed at the same speed, MBH tries to differentiate same actions with varying speed, since it is completely dependent on velocity. But HSGA can recognize two videos of same action correctly, even if they are performed at different speeds. It was experimentally verified that MBH fails terribly in case of camera shakes, but HSGA is more robust to it.

Now we compare our descriptor combination with other descriptor combinations in the literature. Comparison of the recognition accuracy of the descriptor combinations MBH/HOF[25] and HOG/HOF[14] with our proposed descriptor combination is shown in Table 5. All descriptors were computed around improved dense trajectories and were fused using hybrid fusion method. MBH/ HSGA outperforms all other descriptors since the combination encodes the most relevant motion information, *viz.* velocity and acceleration.

Dataset	MBH/HSGA	HOF/HOG[14]	MBH/HOF[25]
YouTube	**84.06%**	82.17%	83.30%
UCF50	**78.21%**	77.74%	77.85%

Table 5. Comparison of different descriptor combinations

Further experiments to evaluate the performance of histogram of optical acceleration (HOA) for action recognition were conducted, but the results did not match up to that of HSGA. Fusing HOA with its velocity counterpart HOF gave a recognition accuracy of 78.87% on YouTube dataset and 71.5% on UCF50 dataset and hence proved to be less effective compared to MBH/ HSGA combination.

6. Abnormal Event Detection

The increase in the volume of surveillance footages has made automated abnormal event detection a greater necessity than before. Anomaly detection aims to find any event deviating from the normal pattern and this section briefs the use of optical acceleration in this scenario. Problem of anomaly detection is not a typical classification problem, since it is not possible to train using all negative samples. Common trend is to learn the normal patterns from training videos and identify any event deviating from normal ones as abnormal. A wide variety of training or learning framework ranging from statistical models like Hidden Markov Models (HMM)[13] to binary classifiers like support vector machines (SVM) [5] were used to detect unusual activities in surveillance videos. Some of the recent works perform abnormal event detection in a sparse reconstruction framework which consist of learning sparse dictionary from normal training data and thresholding sparse reconstruction error to detect anomaly. An important one in this category learns a dictionary from multi level histogram of optical flow features extracted from normal training videos [4]. A number works that concentrated on enhancing the dictionary learning by using sparse combination learning[18], computing similarity measure between neighbouring blocks [1] and learning an additional abnormal dictionary while testing is on the go [29].

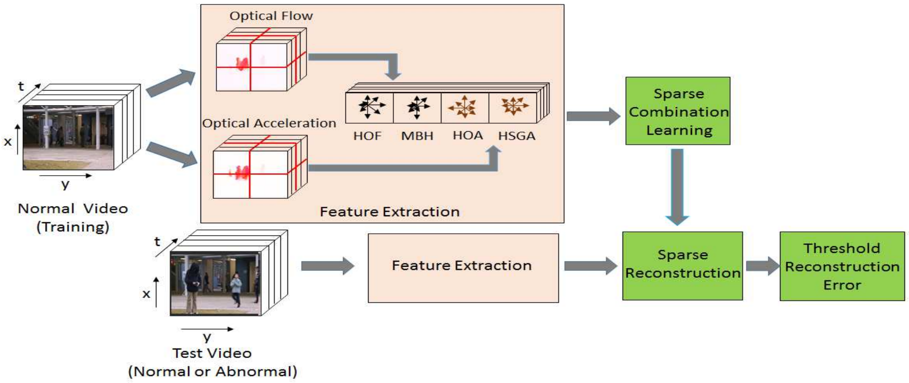

Figure 5. Proposed anomaly detection system

6.1. Optical Acceleration for Anomaly Detection

Figure 5 shows our proposed abnormal event detection which uses a combination of velocity and acceleration descriptors in a sparse reconstruction framework. Each video in the training set is divided into $10 \times 10 \times 5$ patches and for each patch, velocity (HOF and MBH) and acceleration (HOA and HSGA) descriptors are computed and concatenated to form the feature vector. Let the concatenated training feature set extracted from the normal training videos be $X = [f_1, f_2, f_3, ..., f_m,]$, where m is the total number of $10 \times 10 \times 5$ patches in the training video. From this feature set, a normal sparse dictionary set $D = [D_1, D_2, D_3, ..., D_k]$ is learned using sparse combination learning[18]. A normal data can be represented using one of the dictionary vectors from D, but the reconstruction error while representing an abnormal data will be high. A test video is divided into $10 \times 10 \times 5$ patches and for each patch the feature vector f is extracted. We check for the dictionary element that can represent f with minimum error and this error is thresholded to classify the test video as normal or abnormal.

6.2. Experiments and Discussions

We have experimented our method on Avenue dataset[18]. It contains 16 normal videos for training and 21 testing video clips with some abnormal activity captured in Chinese University of Hong Kong (CUHK) campus. A pixel level binary mask is provided as ground truth and hence evaluation is possible in frame level as well as in pixel level. Recognition accuracy is the suggested performance measure for this dataset.

Descriptor	Accuracy
Velocity	64.3%
Acceleration	64.2%
Combined	**68.6%**

Table 6. Recognition accuracy of velocity, acceleration and combined descriptors

Table 6 shows the pixel level recognition accuracy of abnormal event detection using velocity descriptors (HOA and MBH), acceleration descriptors(HOA and HSGA) and the combination of the four. As in case of action recognition, the combination of velocity and acceleration descriptors shows a superior performance.

Performance comparison of our descriptor combination with multilevel HOF (mHOF) and spatio-temporal gradient is presented in table 7. Multilevel HOF is derived from optical flow and encodes only velocity information, while spatio-temporal gradient does not differentiate space and time properties. Our descriptor combination encodes both acceleration velocity, shows an improved performance.

Feature	Avenue dataset
Multilevel HOF[4]	67.1%
3D time gradient[18]	67.3%
Our feature	**68.6**%

Table 7. Comparison with existing descriptors

7. Conclusion

In this paper, we developed algorithms to compute optical acceleration for encoding acceleration of each pixel in a video. Extending this concept we use two acceleration descriptors histogram of optical acceleration (HOA) and histogram of spatial gradient of acceleration (HSGA) for motion description. We have demonstrated the use of acceleration descriptors for efficient recognition of human actions and detection of anomalies in videos. Experimental results show that acceleration is as important as velocity to represent motion information contained in videos. Acceleration descriptors could provide some information that velocity descriptors missed and hence combining them improved motion description, when compared to using either of them alone. The motion descriptors introduced as part of this paper may pave way for more interesting applications in computer vision.

References

[1] S. Biswas and R. V. Babu. Sparse representation based anomaly detection with enhanced local dictionaries. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 5532–5536. IEEE, 2014.

[2] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[3] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *Proceedings of the European conference on computer vision (ECCV)*, pages 282–295. Springer, 2010.

[4] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3449–3456. IEEE, 2011.

[5] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas. Abnormal detection using interaction energy potentials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3161–3167. IEEE, 2011.

[6] A. Dosovitskiy, P. Fischery, E. Ilg, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, T. Brox, et al. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766. IEEE, 2015.

[7] A. Edison and C. Jiji. Hsga: A novel acceleration descriptor for human action recognition. In *Proceedings of the National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, pages 1–4. IEEE, 2015.

[8] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of Scandinavian Conference on Image Analysis*, volume 2749. Springer, 2003.

[9] J. J. Gibson. *The perception of the visual world.* Houghton Mifflin, 1950.

[10] B. K. Horn and B. G. Schunck. Determining optical flow. In *1981 Technical symposium east*, pages 319–331. International Society for Optics and Photonics, 1981.

[11] H. Kataoka, Y. He, S. Shirakabe, and Y. Satoh. Motion representation with acceleration images. In *Proceedings of the ECCV 2016 Workshops*, pages 18–24. Springer, 2016.

[12] A. Klaser, M. Marszałek, C. Schmid, et al. A spatio-temporal descriptor based on 3d-gradients. In *Proceedings of British Machine Vision Conference*, 2008.

[13] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1453. IEEE, 2009.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[15] A. Li, Z. Miao, Y. Cen, and Q. Liang. Abnormal event detection based on sparse reconstruction in crowded scenes. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1786–1790. IEEE, 2016.

[16] J. Li, S. G. Nikolov, C. P. Benton, and N. E. Scott-Samuel. Adaptive summarisation of surveillance video sequences. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS).*, pages 546–551. IEEE, 2007.

[17] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1996–2003, 2009.

[18] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2720–2727, 2013.

[19] H. Nallaivarothayan, C. Fookes, S. Denman, and S. Sridharan. An MRF based abnormal event detection approach using motion and appearance features. In *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 343–348, 2014.

[20] J. L. Prince and E. R. McVeigh. Motion estimation from tagged mr image sequences. *IEEE Transactions on Medical Imaging*, 11(2):238–249, 1992.

[21] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013.

[22] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Proceedings of International Conference on Pattern Recognition*, volume 3, pages 32–36, 2004.

[23] L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black. Optical flow with semantic segmentation and localized layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[24] H. Van Luong, L. L. Raket, X. Huang, and S. Forchhammer. Side information and noise learning for distributed video coding using optical flow and clustering. *IEEE Transactions on Image Processing*, 21(12):4782–4796, 2012.

[25] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013.

[26] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of European Conference on Computer Vision*, pages 650–663. Springer-Verlag, 2008.

[27] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Proceedings of IEEE conf. on Computer Vision and Pattern Recognition*, pages 379–385, 1992.

[28] L. Zhang, Y. Feng, J. Han, and X. Zhen. Realistic human action recognition: When deep learning meets vlad. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1352–1356. IEEE, 2016.

[29] Z. Zhu, J. Wang, and N. Yu. Anomaly detection via 3d-hof and fast double sparse representation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 286–290. IEEE, 2016.