Spatial-temporal Motion Information Integration for Action Detection and Recognition in Non-static Background

Dianting Liu, Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33124, USA
d.liu4@umiami.edu, shyu@miami.edu

Guiru Zhao
China Earthquake Networks Center
5 Nanheng ST, Sanlihe Rd, Xicheng District, Beijing 100045, P. R. China
zgr@seis.ac.cn

Abstract

Various motion detection methods have been proposed in the past decade, but there are seldom attempts to investigate the advantages and disadvantages of different detection mechanisms so that they can complement each other to achieve a better performance. Toward such a demand, this paper proposes a human action detection and recognition framework to bridge the semantic gap between low-level pixel intensity change and the high-level understanding of the meaning of an action. To achieve a robust estimation of the region of action with the complexities of an uncontrolled background, we propose the combination of the optical flow field and Harris3D corner detector to obtain a new spatial-temporal estimation in the video sequences. The action detection method, considering the integrated motion information, works well with the dynamic background and camera motion, and demonstrates the advantage of the proposed method of integrating multiple spatial-temporal cues. Then the local features (SIFT and STIP) extracted from the estimated region of action are used to learn the Universal Background Model (UBM) for the action recognition task. The experimental results on KTH and UCF YouTube Action (UCF11) data sets show that the proposed action detection and recognition framework can not only better estimate the region of action but also achieve better recognition accuracy comparing with the peer work.

Keywords: Spatio-temporal Motion Information Integration, Action Detection, Action Recognition, Universal Background Model (UBM), Gaussian Mixture Models (GMM), GMM Supervector.

1. Introduction

In the recent years, video content analysis and human action recognition have been used in a broad range of applications in real-time surveillance, activity monitoring, video indexing and retrieval, human-computer interaction, etc. [6, 20, 22, 30]. A batch of action detection and recognition models have been proposed and achieved good performance in videos captured under controlled backgrounds (as shown in Figure 1) [15, 23, 27]. Nevertheless, more progresses toward model robustness are expected in order to handle the complexities of unconstrained backgrounds, such as videos recoded by an amateur using a hand-held camera containing significant camera motion, background clutter, and changes in object appearance, scale, and illumination conditions. These uncontrolled videos are the major challenges to the multimedia retrieval engines on the Internet, which call for rapid summarization and processing algorithms. The drawbacks of the most existing techniques include the requirements of (1) static cameras or approximate compensation of camera motion; (2) foreground objects that move in a consistent direction or have faster variations in appearance than the background; and (3) explicit background models [26]. These requirements are generally unrealistic and particularly questionable when an egomotion happens, e.g., a camera that tracks an action in a manner such that the latter has a very small optical flow, or the background is dynamic. In addition, background learning requires a training set of background-only images [33] or batch processing (e.g., median filtering [9]) of a large number of video frames, which must be repeated for each scene and is difficult for dynamic scenes (where the background changes continuously).

Figure 1. Examples of UCF Youtube action (UCF11) data set with approximately 1,160 videos in 11 categories [23]

To overcome the above challenges, we propose a robust action detection and recognition framework that integrates multiple motion detectors and takes the complementary advantages of the motion cues to estimate the region of action. Features extracted from the region are minimally disturbed by scene noise and represent the characteristics of the action. To the best of our knowledge, not much work has been reported on the region detection of action from unconstrained videos in an unsupervised way. One related work is by Liu *et al.* on recognizing actions from videos "in the wild" [23]. They estimated the centroid of the region of action by using the mean of the coordinates of the interest points. The dimensions of the region are calculated by the second central moments of the corresponding centroid. This strategy can obtain good results when the interest points are mainly located on the action, but it would fail when the background is non-static since it contributes a lot of interest points. Ikizler-Cinbis *et al.* [15] estimated the location(s) of the person(s) by using the human detector proposed by Felzenswalb *et al.* [12]. To fill the gap in which the person detector did not fire due to the motion blur and pose variations, the mean-shift tracking method was used to locate the person in every frame [7]. The approach, to some degree, was able to capture the human-related features in the video. However, it is not particularly designed from the perspectives of action detection and recognition and enhance consequently fails to contribute to the action-oriented information. Optical flow is utilized by Reddy *et al.* [27] to give a rough estimate of the velocity at each pixel given two consecutive frames. They, then, applied a threshold on the magnitude of the optical flow to decide if the pixel is moving or stationary. The stationary pixels are regarded as background, while the moving pixels are viewed as the region of action. This method performs well in videos with static scenes, but the strategy fails in the realistic videos with the unconstrained background.

In this paper, we investigate the ideas of motion detectors and propose a framework that detects region(s) of action by integrating multiple spatial-temporal cues and recognizes actions by using static and motion features on the region of action. The main contributions of this paper are summarized as follows.

1. A weighted integration approach is proposed to fuse spatial-temporal information from the optical flow field and the Harris3D detector into a new robust motion representation in the videos.

2. The idea of integral density is utilized to estimate the region of action by using the new motion field. The region of action is defined as the area with a high density of motion.

3. SIFT and STIP features extracted from the region of action are employed to train the universal background model (UBM) for the purpose of action recognition, instead of using the whole feature set. This method is verified to be an effective and efficient way of training recognition model.

To the best of our knowledge, no one has trained UBM by using only action-related features (less than 20% of the whole feature set) and is able to receive a better performance than using the full feature set.

The rest of the paper is organized as follows. Section 2 describes the details of the region of action estimation by integrating multiple spatial-temporal motion fields and quickly locating the high density area of motion. In Section 3, we present the method of action recognition that uses multiple features from the region of action to train UBM and classifies the actions. The experiments and results of the KTH and UCF11 data sets with discussions are provided in Section 4. Finally, a conclusion is given in Section 5.

2. Action Region Estimation by Integrating Spatial-Temporal Motion Information

The state-of-the-art action recognition approaches mainly use the features extracted from the whole frame, no matter the background or the region of action, to generate the code book which inevitably involves unrelated scene information that may affect the recognition performance. In order to decrease the influence of the background on the action recognition task, a new action region estimation method is presented in this section. The proposed algorithm comprehensively analyzes and integrates the motion information on space and time in an unsupervised manner, and is robust to non-static scene and camera motion. The motion features extracted from the estimated region of action later are employed to learn the Universal Background Model (UBM) for the action recognition purposes, which is able to achieve a good performance. The proposed framework is shown in Figure 2.

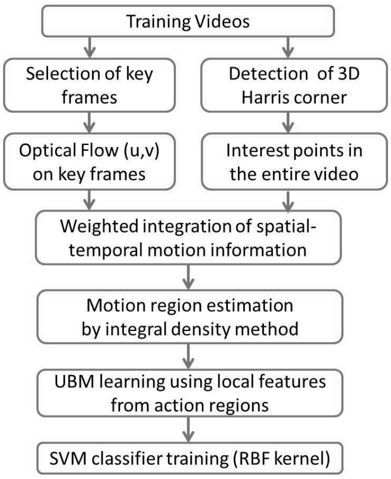

Training Videos
Selection of key frames
Detection of 3D Harris corner
Optical Flow (u,v) on key frames
Interest points in the entire video
Weighted integration of spatial-temporal motion information
Motion region estimation by integral density method
UBM learning using local features from action regions
SVM classifier training (RBF kernel)

Figure 2. Proposed Framework

2.1 Biological Motivation

Psychological studies find that a human vision system perceives external features separately [31] and is sensitive to the difference between the target region and its neighborhood. Such kind of high contrast is more likely to attract human's first sight than their surrounding neighbors [11]. Extensive psychophysics experiments have shown that these mechanisms can be driven by a variety of features, including intensity, color, orientation, or motion, and local feature contrast plays a predominant role in the perception of saliency. Neurophysiological experiments on primates have also shown that neurons in the middle temporal (MT) visual area compute local motion contrast with center-surround mechanisms. In fact, it has been hypothesized that such neurons underlie the perception of motion pop-out and figure-ground segmentation [2]. The center-surround saliency mechanisms of biological systems support the idea of motion region estimation on measurements of local motion contrast. There is no need for training samples or pre-build a "global background model" for the testing instances, which is one of the advantages of the proposed method. Instead, a motion region can be efficiently calculated using merely local motion information and could immediately adapt to different kinds of unknown scenes. Also, using local motion contrast could make the model robust to the camera motion and dynamic background.

2.2 Apparent Motion Descriptor - Optical Flow

Optical flow is the pattern of motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer (an eye or a camera) and the scene. In 1981, Horn and Schunck [1, 14] deduced a basic equation of optical flow estimation when the interval of consecutive frames was short, and the gray change in the image was also small. If at time t, the coordinates of a pixel on the image with its gray value is $I(x, y, t)$, and at time $(t + \triangle t)$, the pixel has moved to new position, its location on the image becomes $(x + \triangle x, y + \triangle y)$, and the gray value becomes $I(x + \triangle x, y + \triangle y, t + \triangle t)$. $dI(x, y, t)/dt = 0$ is got based on the assumption that intensity is conserved. Then the equation can be re-written as $I(x, y, t) = I(x+\triangle x, y+\triangle y, t+\triangle t)$, whose Tayor expansion can derive the gradient constraint equation as below.

$$\tfrac{\partial I}{\partial x}\tfrac{dx}{dt} + \tfrac{\partial I}{\partial y}\tfrac{dy}{dt} + \tfrac{\partial I}{\partial t} = 0.$$

Suppose u and v are two components of the optical flow along the x coordinate and y coordinate, and they are defined as $u = dx/dt$, $v = dy/dt$. Then the basic optical flow equation is obtained as $I_x u + I_y v + I_t = 0$, where I_x denotes the partial x coordinate derivative of $I(x, y, t)$, I_y denotes the partial y coordinate derivative of $I(x, y, t)$, and I_t denotes the partial time derivative of $I(x, y, t)$.

The advantage of using the optical flow is that it does not require any priori knowledge on the object appearance which satisfies the requirement of an unsupervised method in this paper. The disadvantage is that the computation is usually too complex to be used in real-time applications if there is no special hardware support. With the attempt to reduce the computation complexity of the optical flow technique, the motion vector idea using the optical flow technique to work on the block-level instead of pixel-level motion is adopted.

Motion vector is an integral part of many video compression algorithms which are used for motion compensation. The idea behind block matching is to divide the current frame into a matrix of blocks that are then compared with the corresponding block and its neighbors in the previous frame to determine a motion vector that estimates the movement of a block from one frame to another. For fast motion estimation purposes, we employ the optical flow method to describe the spatial motion of blocks in the frame.

2.3 Harris3D Corner Detector

If the video sequences are captured by a moving camera or in a non-static background, no satisfactory results can be obtained by simply relying on the motion described by optical flow to estimate the action region. Thus, in our proposed framework, the space time interest point detector, Harris3D corner detector [18], is employed to integrate the motion presented by optical flow. The Harris3D corner detector is used to detect the spatial-temporal corners with velocity changes over a sequence of frames.

We consider a 3D window about a space-time point $I(x, y, t)$ and analyze the average intensity change (gradient) as the window is shifted by a small amount (σ, τ) in spatial as well as temporal dimensions (σ is the spatial scale and τ is the temporal scale). The space-time gradient is obtained as $\nabla L = (L_x, L_y, L_t)^T$. The interest point is identified by evaluating the distribution of ∇L within a local neighborhood. The matrix μ of the second moments (a 3-by-3 matrix composed of the first order spatial and temporal derivatives being averaged using a Gaussian weighting function $g(\cdot; \sigma_i^2, \tau_i^2)$) that measures the variation of the gradients. A high variation of ∇L implies large eigenvalues of μ, and the spatial-temporal corners are obtained from the local maxima of H over $I(x, y, t)$. That is,

$$H = det(\mu) - k \cdot trace^3(\mu) = \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3,$$

where λs are the eigenvalues of H and k is a constant with a value close to 0.15.

2.4. Integrated Spatial-Temporal Motion

The above discussion shows that the optical flow field and Harris3D corner detector have their individual characteristics in the spatial-temporal motion calculation. The integration of these two sources of motion information may provide the complementary motion information to improve the region of action estimation.

Suppose N key frames are sampled from an action video sequence, and $N - 1$ optical flow fields are generated. Spatial-temporal volumes created around the Harris3D corners are illustrated in gray boxes in Figure 3. All volumes are clustered into $N - 1$ groups based on the time stamp of the key frames. As shown in Figure 3, if the center of the volume is between $[n - 0.5, n + 1.5]$, the volume belongs to group (n, $n + 1$). The histogram of Harris3D volumes of group (n, $n + 1$) is then generated based on the distribution of the volumes along the time line. The new motion $M(x, y)$ at pixel (x, y) between key frames n and $n + 1$ is calculated as given in Equation (1).

$$M(x, y) = O(x, y) * H(x, y) \qquad (1)$$

where $O(x, y)$ is the the motion vector of optical flow and $H(x, y)$ is the histogram of Harris3D volumes at pixel (x,y). This method is also viewed as a weighted optical flow approach which uses the histogram of Harris3D volumes to weigh the optical flow field. In this way, two sources of motion information are integrated in terms of the key frames.

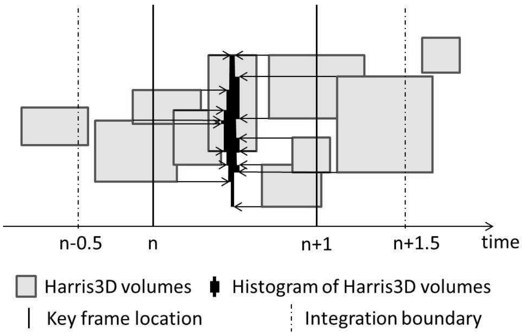


Figure 3. Illustration of the histogram of Harris3D volumes

2.5. Region of Action Estimation

An unsupervised action region estimation method is proposed in this paper by analyzing the new motion field generated from the optical flow and the histogram of Harris3D volumes. The idea of an integral density, as defined in [21], is adopted since it allows fast implementation of the box type convolution filters. The entry of a summed area table $I_{\sum(\mathbf{x})}$ at a location $\mathbf{x}$=(x,y) represents the sum of all values in the input $2D$ matrix I of a rectangular region formed by the point $\mathbf{x}$ and the origin, i.e.,

$$I_{\sum(\mathbf{x})} = \sum_{i=0}^{i\leq x} \sum_{j=0}^{j\leq y} I(i,j).$$

With $I_{\sum(\mathbf{x})}$ calculated, it takes only four additions to calculate the sum of the values over any upright rectangular areas, independent of their sizes. In the same way, the maximal motion region is identified as the region of action. We define the maximal motion region as an area having the highest motion density as shown below, where $v(x, y)$ indicates the integrated motion at pixel (x, y).

$$\arg\max_R \int\int_R v(x, y)dxdy.$$

3. Action Recognition Framework

Gaussian Mixture Models (GMM) are employed in our proposed framework, whose probability density function (pdf) is given by $p(x|\theta) = \sum_{k=1}^{K} \omega_k N(x|\mu_k, \sum_k)$, where K is the number of Gaussian mixtures, and $\theta = \{\omega_k, \mu_k, \sum_k\}_{k=1}^{K}$ is a set of parameters including a mixing coefficient ω_k and a pdf of Gaussian distribution $N(x|\mu_k, \sum_k)$ with the mean vector μ_k and the variance matrix $\sum_k$. The GMM parameters are estimated by using an expectation maximization (EM) algorithm. The EM algorithm is known as a method for finding the maximum likelihood estimators of a model with latent variables.

SIFT [25] and STIP [18] features are used to describe the action video sequences in the action recognition. However, the number of features (SIFT or STIP) from each video is not enough to estimate the GMM parameters precisely. Thus, we first learn a global GMM (called universal background model (UBM)) by using the features from all training videos, then adapt the UBM parameters in order to fit each particular data distribution. This adaptation is made by using the Maximum A Posteriori (MAP) approach [28]. The first step consists of determining the probabilistic alignment of the training vectors with the UBM Gaussian components. For a Gaussian i in the UBM, we compute:

$$\begin{aligned} Pr(i, x_t) &= \frac{\omega_i p_i(x_t)}{\sum_{j=1}^{M} \omega_j p_j(x_t)}; \\ n_i &= \sum_{t=1}^{T} Pr(i, x_t); \\ E_i(x) &= \frac{1}{n_i} \sum_{t=1}^{T} Pr(i, x_t) x_t. \end{aligned}$$

Here, X_t represents the tth feature vector of the video to be modeled. These statistical values are then used for adapting the mean vector $\hat{\mu}$ of each Gaussian.

$$\begin{aligned} \hat{\mu_i} &= \alpha_i E_i(x) + (1 - \alpha_i)\mu_i; \\ \alpha_i &= \frac{n_i}{n_i + r}, \end{aligned}$$

where r is a fixed "relevance factor", usually set between 8 and 20 [5]. The concatenation of all the mean vectors of the N Gaussian components is called the GMM supervector which is first proposed as a speaker recognition method [3] and then has been applied to semantic indexing [16] and music similarity [5]. Knowing the parameter of the UBM, a particular video model can be resumed by the mean vectors of its Gaussian mixture components. The testing videos are classified by using the support vector machines (SVMs) with the RBF kernel [8]. In this paper, to save UBM training time, only features extracted from each region of action are used to model the overall data distribution.

4. Experiments

The detection and recognition experiments were conducted on the KTH and UCF Youtube Action (UCF11) data sets. The KTH data set has 25 actors performing six actions four times in four different environments, resulting in 599 video sequences in total. The video sequences were recorded in a controlled setting with slight camera motion and a simple background. The six categories of actions are boxing, hand clapping, hand waving, jogging, walking, and running [29]. UCF Youtube Action (UCF11) data set is more challenging than the KTH data set since it includes 1,160 videos and has 11 categories of actions collected from YouTube with the non-static background, low quality, camera motions, poor illumination conditions, etc. [23].

4.1. Experimental Setup

For the interest point detection, the Difference of Gaussian (DoG) edge detection method proposed by Lowe [25] and the Harris3D corner detector proposed by Laptev [18] are used to locate the interest points of SIFT and STIP, respectively. Three key frames equally sampled along the video for SIFT feature extraction and optical flow computation. The dimensions of the SIFT and STIP features are reduced to 32 by applying the Principle Component Analysis [17] from 128 and 162, respectively. The number of Gaussian components in GMM (i.e., K) is set to 256 for the KTH and UCF11 data sets. SVM is used to cope with the multi-class classification task. We adopt the empirical setting in libSVM [4], and for comparison purposes, the leave one out cross validation (LOOCV) scheme is employed to compare with some existing approaches.

4.2. Experimental Results on the KTH Data Set

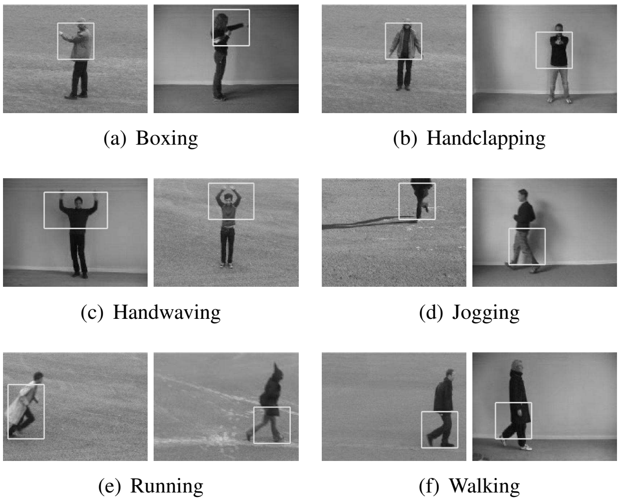
Though the proposed framework is mainly designed to deal with videos captured in unconstrained environments, it is also proved to achieve pretty good performance in videos recorded in a "clean" background, such as the KTH data set. First, the accurate localization of an action is verified. Sample regions of action estimation results are illustrated in Figure 4. The features extracted from the regions of action were used to learn UBM and a classification accuracy of 93.67% is obtained if combining the SIFT and STIP similarity scores, whereas the accuracy is 84.65% if using the SIFT features alone and is 90.65% if using the STIP features alone. The combination of two kinds of features achieves 3% improvement in the performance. Table 1 lists several state-of-the-art performance results on the KTH data set, and indicates that our proposed framework outperforms the peer work. Please note that the amount of features used to train UBM is less than 15% of the total features over all video sequences, which clearly shows to reduce lots of offline training time.

Figure 5 shows the confusion table containing the detailed confusion values between action categories. Based

Table 1. Accuracy comparison on the KTH data set (%)

Algorithm	Accuracy (%)
Proposed framework	**93.67**
Reddy *et al.* [27]	89.79
Dollar *et al.* [10]	81.2
Liu *et al.* [24]	91.3
Wong *et al.* [32]	83.9
Laptev *et al.* [19]	91.8

on the moving part of a person, the six action categories can be grouped into the limb action (boxing, hand clapping, and hand waving) and leg action (jogging, running, and walking). Please note that the confusion happens either within limb action or leg action videos. From the figure, it can be seen that no limb action is misclassified as leg action, and vice versa. This indicates our proposed framework is reasonably good.



(a) Boxing

(b) Handclapping

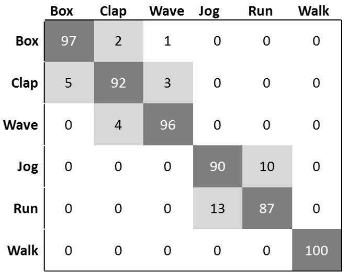
(c) Handwaving

(d) Jogging

(e) Running

(f) Walking

Figure 4. Region of action detection results on sample frames in the KTH data set.

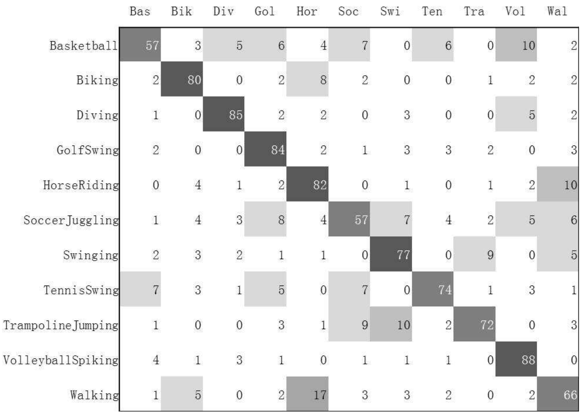


	Box	Clap	Wave	Jog	Run	Walk
Box	97	2	1	0	0	0
Clap	5	92	3	0	0	0
Wave	0	4	96	0	0	0
Jog	0	0	0	90	10	0
Run	0	0	0	13	87	0
Walk	0	0	0	0	0	100

Figure 5. Confusion matrix of 6 action categories on the KTH data set with an average performance of 93.67%.

4.3. Experimental Results on the UCF11 Data Set

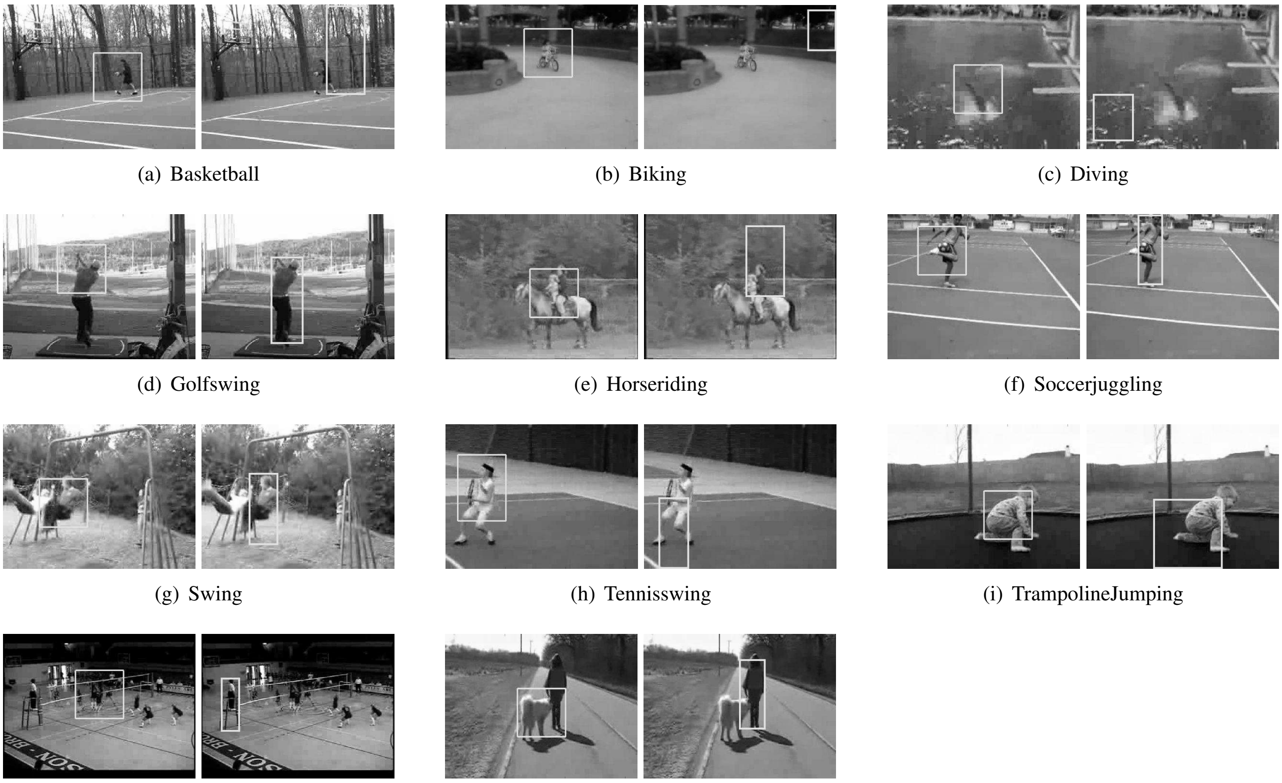
The UCF11 data set is more challenging than the KTH data set, since it contains realistic actions, camera motions, and complicated backgrounds. Figure 6 illustrates some sample results of motion region estimation of the proposed framework (on the left of each sub-figure) and felzenszwalb's part-based models (on the right of each sub-figure) [12]. The codes we used to conduct felzenszwalb's algorithm was downloaded from [13]. Note that felzenszwalb's method works well with human vertical positions in simple backgrounds, such as in Figures 6(d), 6(f) and 6(k). Since the method does not consider temporal information, it may fail in cluttered scenes such as in Figure 6(a) which has a lot of trees having similar appearances with the person. In contrast, since our proposed framework is unsupervised, it could effectively locate the region of action without many appearance constraints obtained from the training data. Furthermore, our proposed framework is motion-driven so it is more suitable for action detection which includes the interaction of humans and objects like biking (Figure 6(b)), horse riding (Figure 6(e)), etc.



	Bas	Bik	Div	Gol	Hor	Soc	Swi	Ten	Tra	Vol	Wal
Basketball	57	3	5	6	4	7	0	6	0	10	2
Biking	2	80	0	2	8	2	0	0	1	2	2
Diving	1	0	85	2	2	0	3	0	0	5	2
GolfSwing	2	0	0	84	2	1	3	3	2	0	3
HorseRiding	0	4	1	2	82	0	1	0	1	2	10
SoccerJuggling	1	4	3	8	4	57	7	4	2	5	6
Swinging	2	3	2	1	1	0	77	0	9	0	5
TennisSwing	7	3	1	5	0	7	0	74	1	3	1
TrampolineJumping	1	0	0	3	1	9	10	2	72	0	3
VolleyballSpiking	4	1	3	1	0	1	1	1	0	88	0
Walking	1	5	0	2	17	3	3	2	0	2	66

Figure 7. Confusion matrix of 11 action categories on the UCF data set with an average performance of 76.06%.

In addition, unlike previous approaches that use all features in the videos (extracted from the scene and object), we use only those features extracted from the region of action to train UBM, which significantly reduces the training time. In this experiment, the action-related features are about 20% of the full feature set (scene and object), but our proposed framework achieves better performance than the previous



(a) Basketball (b) Biking (c) Diving

(d) Golfswing (e) Horseriding (f) Soccerjuggling

(g) Swing (h) Tennisswing (i) TrampolineJumping

(j) VolleyballSpiking (k) Walkingwithdog

Figure 6. Sample results of the proposed region of action detection method (left) and felzenszwalb's part-based models [12] (right) in UCF11 data set.

approaches which use a full set of features. Please note that our proposed framework achieves the performance of 76.06% (as presented in Figure 7) after fusing the similarity scores of SIFT and STIP; while the performance obtained from the SIFT descriptor alone is 55.85% and that from the STIP descriptor alone is 72.82%. In our proposed framework, the size of the code book used in the experiments is only 256, which is relatively smaller than those in the state-of-the-art work, and it also achieves good performances. This demonstrates that the features are extracted from the correct regions of actions and can describe the class-related information. The recognition performance reported by Liu $et\ al.$ is 71.2% using hybrid features obtained by pruning the motion and static features [23]. Another similar work that split the moving foreground from the static background and then combined the motion and the scene context features obtained 73.2% [27].

5 Conclusions

In this paper, a new action detection and recognition framework that integrates the spatial-temporal motion obtained from the optical flow field and the Harris3D corner detector is proposed. It is motivated by taking the advantages of the two sources of motion information identified by different methods to obtain the complementary motion information which is kept in the new motion representation. A fast region of action estimation method is also proposed by using the integral density algorithm. The SIFT and STIP features extracted from the regions are used to learn UBM for the action recognition proposes. The experimental results verify that the proposed framework achieves good performance on both action detection and recognition tasks.

Acknowledgments

For Guiru Zhao, this work was financially supported in part by the program of China Scholarships Council (No. 201204190040) and the program of research on visualization of seismic monitoring information and integrated service technology (No. 201208010).

References

[1] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.

[2] R. Born, J. M. Groh, R. Zhao, and S. J. Lukasewycz. Segregation of object and background motion in visual area mt: Effects of microstimulation on eye movements. *Neuron*, 26:725–734, 2000.

[3] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308–311, 2006.

[4] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[5] C. Charbuillet, D. Tardieu, and G. Peeters. Gmm-supervector for content based music similarity. In *Proceedings of the International Conference on Digital Audio Effects*, pages 425–428, 2011.

[6] S.-C. Chen, M.-L. Shyu, S. Peeta, and C. Zhang. Spatiotemporal vehicle tracking: The use of unsupervised learning-based segmentation and object tracking. *IEEE Robotics and Automation Magazine*, 12(1):50–58, March 2005.

[7] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

[8] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[9] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, 2003.

[10] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72. IEEE, 2005.

[11] J. Duncan and G. Humphreys. Visual search and stimulus similarity. *Psychological Review*, 96(3):433–58, July 1989.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[13] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/.

[14] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1):185–203, 1981.

[15] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *Computer Vision*, pages 494–507. Springer, 2010.

[16] N. Inoue and K. Shinoda. A fast map adaptation technique for gmm-supervector-based video semantic indexing systems. In *Proceedings of the 19th ACM international Conference on Multimedia*, pages 1357–1360. ACM, 2011.

[17] I. T. Jolliffe. *Principal component analysis*, volume 487. Springer-Verlag New York, 1986.

[18] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

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

[20] L. Lin, C. Chen, M.-L. Shyu, and S.-C. Chen. Weighted subspace filtering and ranking algorithms for video concept retrieval. *IEEE Multimedia*, 18(3):32–43, 2011.

[21] D. Liu and M.-L. Shyu. Effective moving object detection and retrieval via integrating spatial-temporal multimedia information. In *Proceedings of the IEEE International Symposium on Multimedia*, pages 364–371. IEEE, 2012.

[22] D. Liu, M.-L. Shyu, Q. Zhu, and S.-C. Chen. Moving object detection under object occlusion situations in video sequences. In *Proceedings of the IEEE International Symposium on Multimedia*, pages 271–278, 2011.

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

[24] J. Liu and M. Shah. Learning human actions via information maximization. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[25] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60(2):91–110, Nov. 2004.

[26] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):171–177, 2010.

[27] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, pages 1–11, 2012.

[28] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1):19–41, 2000.

[29] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th IEEE International Conference on Pattern Recognition*, volume 3, pages 32–36. IEEE, 2004.

[30] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen. Video semantic event/concept detection using a subspace-based multimedia data mining framework. *IEEE Transactions on Multimedia*, 10(2):252–259, February 2008.

[31] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, January 1980.

[32] S.-F. Wong, T.-K. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–6. IEEE, 2007.

[33] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 2, pages 28–31. IEEE, 2004.