Machine Vision and Applications MVAP-D-12-00244

Classifying Web Videos using a Global Video Descriptor

Berkan Solmaz · Shayan Modiri Assari · Mubarak Shah

Received: 24 July 2012 / Revised: 28 August 2012 /Accepted: 31 August 2012

Abstract Computing descriptors for videos is a crucial task in computer vision. In this paper, we propose a global video descriptor for classification of videos. Our method, bypasses the detection of interest points, the extraction of local video descriptors and the quantization of descriptors into a code book; it represents each video sequence as a single feature vector. Our global descriptor is computed by applying a bank of 3-D spatio-temporal filters on the frequency spectrum of a video sequence; hence it integrates the information about the motion and scene structure. We tested our approach on three datasets, KTH[1], UCF50 [2] and HMDB51[3], and obtained promising results which demonstrate the robustness and the discriminative power of our global video descriptor for classifying videos of various actions. In addition, the combination of our global descriptor and a local descriptor resulted in the highest classification accuracies on UCF50 and HMDB51 datasets.

Keywords Video descriptors · Action recognition · Frequency spectrum · Spatio-temporal analysis

B. Solmaz
4000 Central Florida Blvd, Orlando, FL 32816
Tel.: +1-407-8234733
E-mail: bsolmaz@eecs.ucf.edu

S. Modiri Assari
4000 Central Florida Blvd, Orlando, FL 32816
Tel.: +1-407-8234733
E-mail: smodiri@eecs.ucf.edu

M. Shah
4000 Central Florida Blvd, Orlando, FL 32816
Tel.: +1-407-8235077
Fax: +1-407-8230594
E-mail: shah@eecs.ucf.edu

1 Introduction

Due to the massive number of videos uploaded online each day, recognizing actions and scenes in videos is an important problem in computer vision research. In the literature, several approaches have been proposed for the recognition of actions in diverse real-world videos. Recent survey papers [4,5] provide a detailed overview of the present approaches, challenges and the available datasets.

Laptev et al. [6], addressing the demand for datasets of large number of realistic action classes, introduced Hollywood dataset with eight action classes, collected from the movies. Liu et al. [7] introduced the UCF YouTube dataset consisting of 11 categories of actions collected from YouTube and personal videos. UCF50 [2] extended the 11 action categories of the UCF YouTube dataset to a total of 50 action categories. Kuehne et al. [3] presented the very challenging HMDB51 dataset of 51 action classes with video clips collected from a wide range of sources. Provided these datasets, several approaches of human action recognition that fall into two broad groups were proposed: global methods and local spatio-temporal interest point based methods.

Global methods represent the actions based on holistic information about the action and scene. These methods often require the localization of the human body through alignment, background subtraction or tracking. Some common representations of holistic methods are silhouettes [8,9] and motion [10,11]. Holistic models perform well in a controlled environment; however, on datasets of low-quality web videos, conditions such as the presence of clutter in the background, occlusion, variance in illumination and viewpoint changes make the use of these methods impractical. Recently, trajectory-based methods were proposed for action recog-

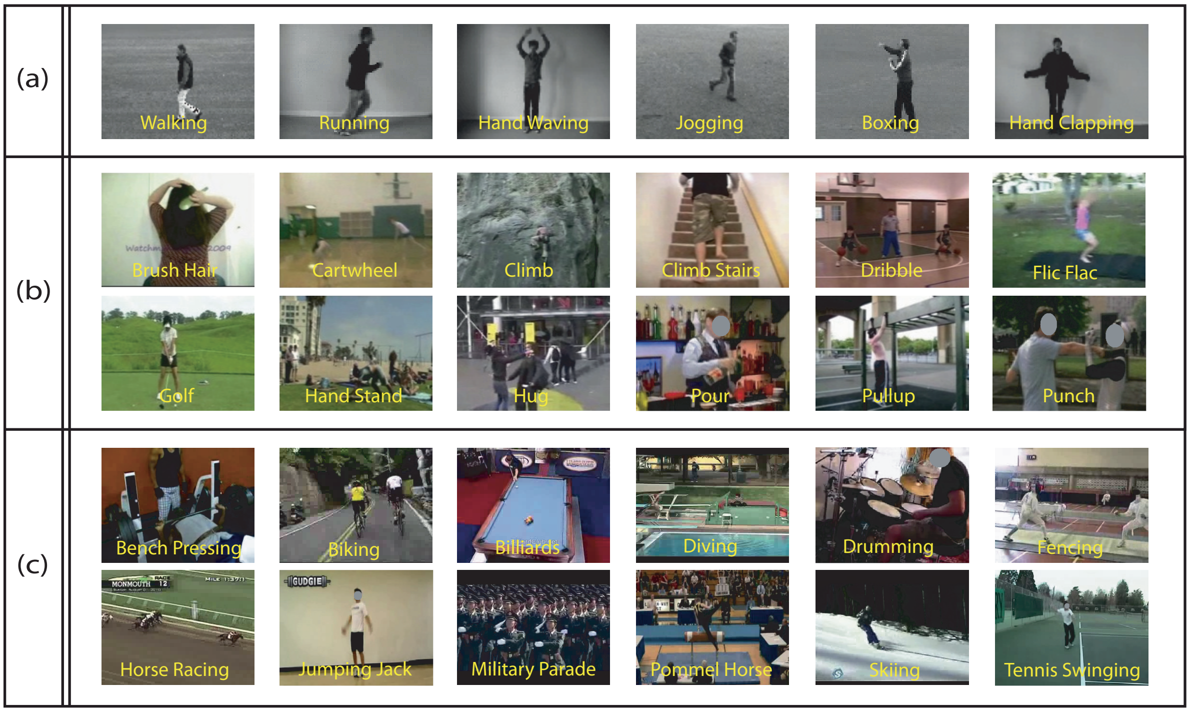


Fig. 1 Example action classes from the **a** KTH, **b** UCF50 and **c** HMDB51 datasets

nition. Wu [12] captured ensemble motions of a scene on a set of dense Lagrangian particle trajectories, which were computed by numerical integration of optical flow over time. Then actions were described by the use chaotic features extracted on object motion trajectories which were obtained after low rank minimization and clustering of all trajectories. Similarly, Wang et al. [13] represented the videos by computing motion boundary histograms, HOG, HOF and trajectory descriptors along the dense foreground motion trajectories. Then, they utilized the standard bag-of-features model for the final representation of the videos. Both of these methods only extracted features on the dynamic trajectories belonging to the scene foreground and did not capture the context information such as the objects present in the background or scene properties, which may be helpful for recognizing actions. In addition, computation and analysis of trajectories may be computationally expensive.

Spatio-temporal interest point-based methods represent the scene and the performed actions as a combination of local descriptors, which are computed in a neighborhood of interest points. The neighborhood can be selected as an image patch or as a spatio-temporal volume, called cuboids, in a video. The spatio-temporal interest point based methods have received a lot of attention in the vision community due to their robustness to scale and viewpoint invariance. Laptev et al. [14] introduced the Space−Time Interest Point detector, a three-dimensional (3-D) variant of the Harris corner detector [15], which identified the points with high variations in intensity and motion. They used a bag-of-features representation on histogram of oriented gradients (HOG) and histogram of optical flow (HOF) descriptors for recognizing natural human actions [6]. Observing the sparseness of the detected STIP interest points, Dollar [16] proposed an alternative feature detector, which computes the response after the application of separable Gaussian filters in space and a quadrature pair of Gabor filters in the time domain for each pixel, followed by the computation of local maxima. Depending on the response, they simply compute gradients or optical flow on cuboids, then flatten them and finally apply principal component analysis (PCA) for dimension reduction. Klaser et al. [17] presented a video descriptor which is based on histograms of oriented 3-D spatio-temporal gradients. Scovanner et al. [18] proposed 3-D SIFT, an extension of the SIFT descriptor to spatio-temporal data. The local descriptor approaches are less sensitive to noise or occlusion; however, they require the detection of sufficient and relevant interest points and lack the capability of modeling the global geometrical or temporal information. Furthermore, these approaches, often utilize the bag-of-features model, requiring the quantization of large amount of data. Even though the interest points and the features are computed locally, each sequence is rep-

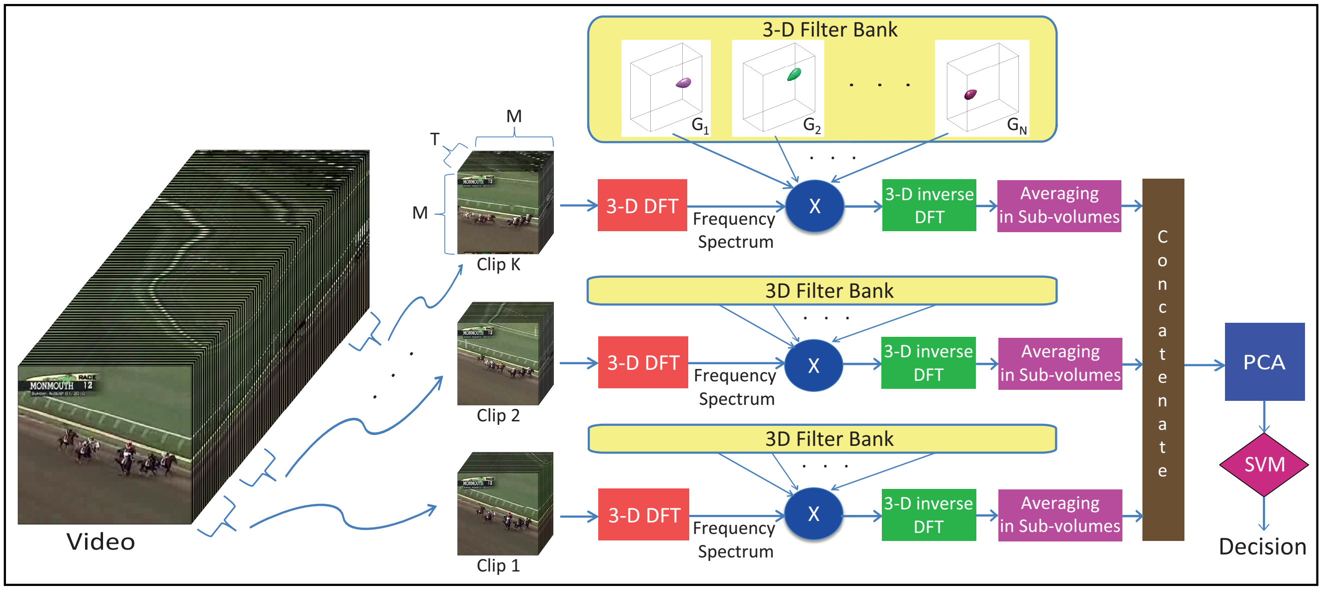


Fig. 2 Overview of our approach: given a video, we extract K clips and compute the 3-D Discrete Fourier Transform. Applying each filter of the 3-D filter bank separately to the frequency spectrum, we quantize the output in fixed sub-volumes. Next, we concatenate the outputs and perform dimension reduction by Principal Component Analysis and classification by the use of a support vector machine

resented by a global histogram, which does not carry any spatial or temporal information.

There are also recent works aimed at capturing the relationships between actions and the scene. Ikizler-Cinbis et al. [19] extracted multiple features on the human, objects and scene, and utilized a multiple-instance learning framework for human action recognition on YouTube videos. However, their approach required motion compensation for foreground estimation, and also the detection and tracking of the human in the scene. Liu et al. [7] extracted a combination of motion and static features and utilized the PageRank algorithm to prune the static features using motion cues as an alternative way to motion compensation. The hybrid use of motion and static features improved the performance of their approach. In our work, we aimed to design a descriptor which captures both motion and scene information without the need of motion compensation.

The "gist" proposed by Torralba et al. [20] is a global scene descriptor based on power spectrum features, and it has the state-of-the-art performance for scene classification. However, it is not suitable for action recognition as it does not capture the motion information. We believe the computation of frequency spectral components of videos may provide useful scene and motion information for action classification. In this paper, we propose a global descriptor for videos, to be used for the action classification of challenging datasets such as UCF50 and HMDB51 with a large number of action classes, some of which are illustrated in Fig. 1. Our descriptor is generated by applying a bank of 3-D spatio-temporal filters on the frequency spectrum of a sequence. The bandpass nature of these filters alleviates the need for motion compensation. Furthermore, as opposed to the approaches which apply bag-of-features model, our approach preserves the spatial and temporal information, as we perform quantization in fixed spatio-temporal sub-volumes after application of each filter on the frequency spectrum and taking the inverse Fourier transform. As the filter responses for all filters on all sub-volumes are concatenated, the ordering and the length of each feature vector are identical for all represented video clips. The framework of our approach is shown in Fig. 2.

In summary, our main contribution in this paper is the presentation of a new global motion and scene descriptor for the classification of realistic videos. We compared the performance of a state-of-the-art local descriptor to our global descriptor. Our global descriptor obtained the highest classification accuracies on two of the most complex datasets, UCF50 and HMDB51, among all published results, to the best of our knowledge. Moreover, the combination of these local and global descriptors resulted in a further improvement in the results.

The structure of our paper is as follows. In Sect. 2 , we present the basics of our approach. We describe the implementation details in Sect. 3. In Sect. 4, we present the quantitative results on the KTH, UCF50

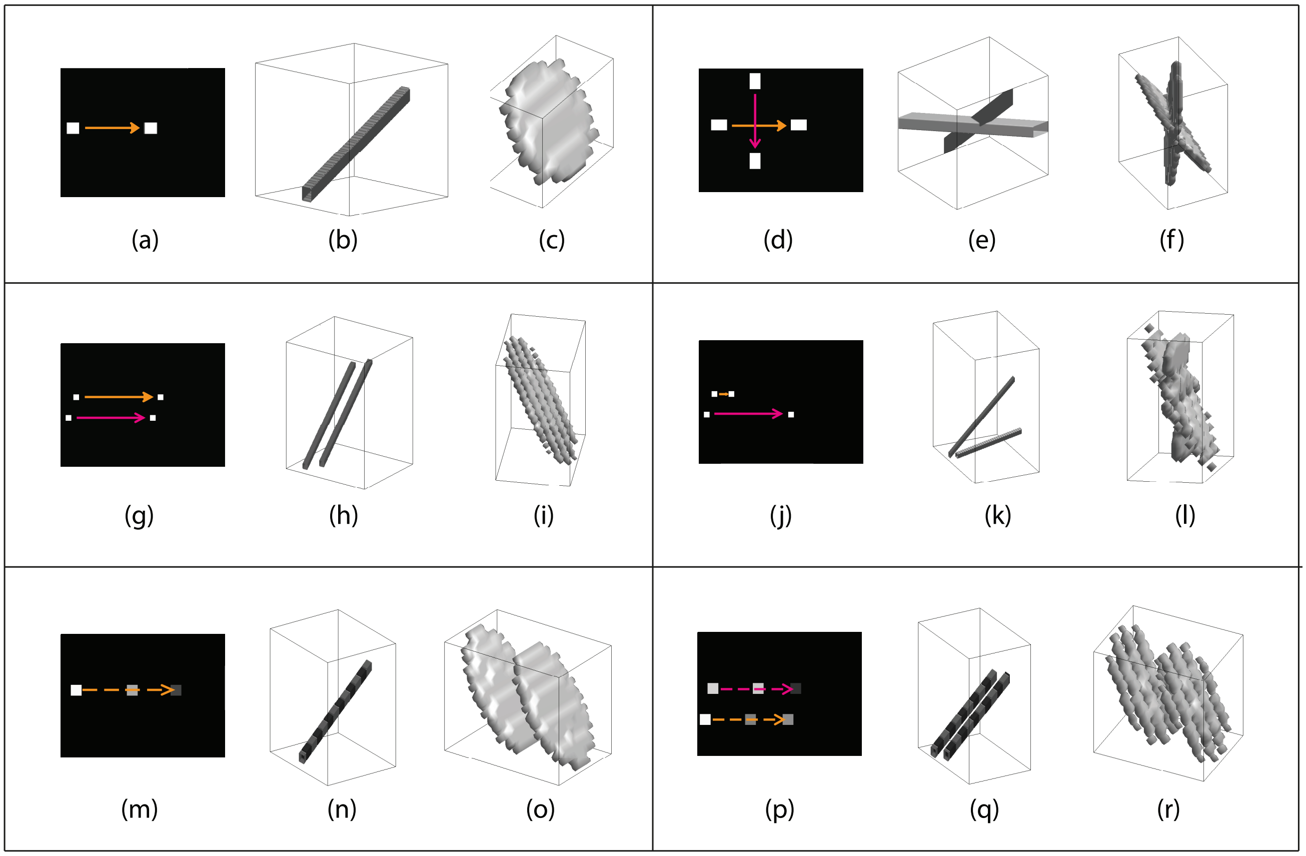


Fig. 3 Orientation of frequency spectrums: the translating object **(a)** generates a space−time volume **(b)**, and a frequency spectrum of non-zero values on a plane **(c)**. Similarly, motion in different orientations **(d)** results in the volume in **e** and frequency spectrum **(f)** with two planes. Uni-directional motion results in a single plane in the frequency spectrum **(g−i)**. Motion with different velocities **(j)** corresponds to two planes in the frequency spectum **(l)**. A translating object with a sinusoidal intensity over time **(m−n)** resulted in two identical planes in frequency spectrum with a separation based on the frequency of the object **(o)**. For multiple objects introducing more gradients **(p, g)**, the planes are still present but appear partially **(i, r)**

and HMDB51 datasets. Finally, we conclude our work in Sect. 5.

2 Gist of a Video

Feature extraction and the computation of descriptors are crucial tasks in action recognition and the classification of videos. In this paper, we introduce a 3-D global descriptor for real-world videos, such as those included in the UCF50 and HMDB51 datasets. Videos which involve similar actions tend to have similar scene structure and motion. The regularities in the appearance or motion can be used to pinpoint the type of actions involved in the video, and can be useful in the classification of videos. The frequency spectrum computed for a video clip could capture both scene and motion information effectively [20–22], as it represents the signal as a sum of many individual frequency components. In a video clip, the frequency spectrum can be estimated by computing the 3-D discrete Fourier transform (DFT).

The motion is an important element which can be representative of the type of performed action in a scene. It can be explained in a straightforward way by considering the problem in the Fourier domain [22]. The frequency spectrum of a two-dimensional pattern translating on an image plane lies on a plane, the orientation of which depends on the velocity of the pattern. Given a 2-D image $f_0(x, y)$, we can create a volume, space−time image sequence, by translating $f_0(x, y)$ with a velocity $\bar{u} = [u_1 u_2]$ over time. This volume is then expressed as

$$f(x, y, t) = f_0(x - u_1 t, y - u_2 t). \tag{1}$$

The three-dimensional Fourier transform of $f(x, y, t)$ over space and time is computed as

$$F(f_x, f_y, f_t) = \frac{1}{MNT} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \sum_{t=0}^{T-1} f(x, y, t) \; e^{-j2\Pi(\frac{xf_x}{M} + \frac{yf_y}{N} + \frac{tf_t}{T})}, \tag{2}$$

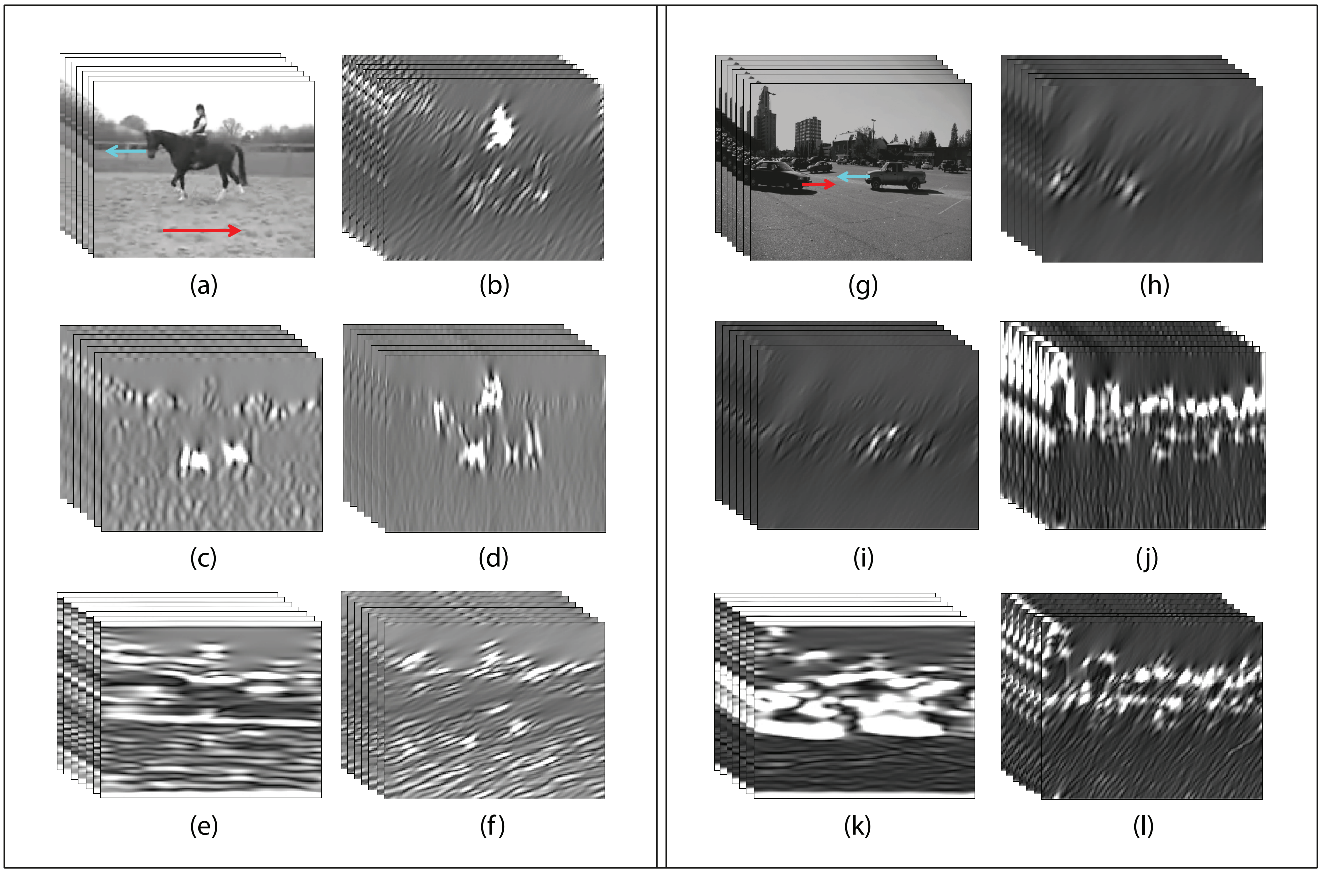


Fig. 4 Effect of filtering the frequency spectrum: using different orientations of 3-D filters on the frequency spectrum for the sample clips **(a, g)**, the components with different motion **(b, c, h, i)**, vertical scene components **(d, j)**, horizontal scene components **(e, k)**, and diagonal scene components **(f, l)** are highlighted. The *red* and *cyan arrows* show the direction of motion in the two videos

where M, N and T are the width, height and length of the clip, and x, y and t are the spatial positions and time of each point in the created volume. Here, the 3-D DFT of the volume will have the same size as the volume itself. After substituting Eq. 1 in Eq. 2 and rearranging the terms, the Fourier transform formula would be

$$F(f_x, f_y, f_t) = \frac{1}{MNT}\sum_{t=0}^{T-1}\left[\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f_0(x-u_1t, y-u_2t) e^{-j2\Pi(\frac{xf_x}{M}+\frac{yf_y}{N})}\right] e^{-j2\Pi\frac{tf_t}{T}}. \quad (3)$$

The inner term in Eq. 3 is actually the 2-D Fourier transform of $f_0(x-u_1t, y-u_2t)$, hence the equation may be simplified to

$$\begin{aligned} F(f_x, f_y, f_t) &= \frac{1}{T}\sum_{t=0}^{T-1} F_0(f_x, f_y) e^{-j2\Pi(\frac{u_1tf_x}{M}+\frac{u_2tf_y}{N})} e^{-j2\Pi\frac{tf_t}{T}}, \\ &= F_0(f_x, f_y)\frac{1}{T}\sum_{t=0}^{T-1} e^{-j2\Pi t(\frac{u_1f_x}{M}+\frac{u_2f_y}{N})} e^{-j2\Pi\frac{tf_t}{T}}, \quad (4) \end{aligned}$$

where $F_0(f_x, f_y)$ represents the 2-D Fourier transform of $f_0(x, y)$. The Fourier transform of the complex exponential term is a Dirac delta function, hence we obtain

$$F(f_x, f_y, f_t) = F_0(f_x, f_y)\,\delta\left(\frac{u_1f_xT}{M}+\frac{u_2f_yT}{N}+f_t\right), \quad (5)$$

where δ is the Dirac delta function. Thus $F(f_x, f_y, f_t)$ will have non-zero values on a plane passing through the origin, as the delta function will be non-zero only when $\left(\frac{u_1f_xT}{M}+\frac{u_2f_yT}{N}+f_t\right) = 0$, as shown in Fig. 3a−c. This derivation shows that analyzing the Fourier transform of a signal, the motion in a sequence can be estimated by finding the plane which contains the power. Furthermore, multiple objects with different motion will generate frequency components in multiple planes as depicted in Fig. 3d−r.

Since the motion can occur in different directions and frequencies, in our work we use 3-D Gabor filters of different orientations and center frequencies to effectively capture the motion information in a video clip. By filtering the frequency spectrum with a certain oriented filter and taking the inverse Fourier transform,

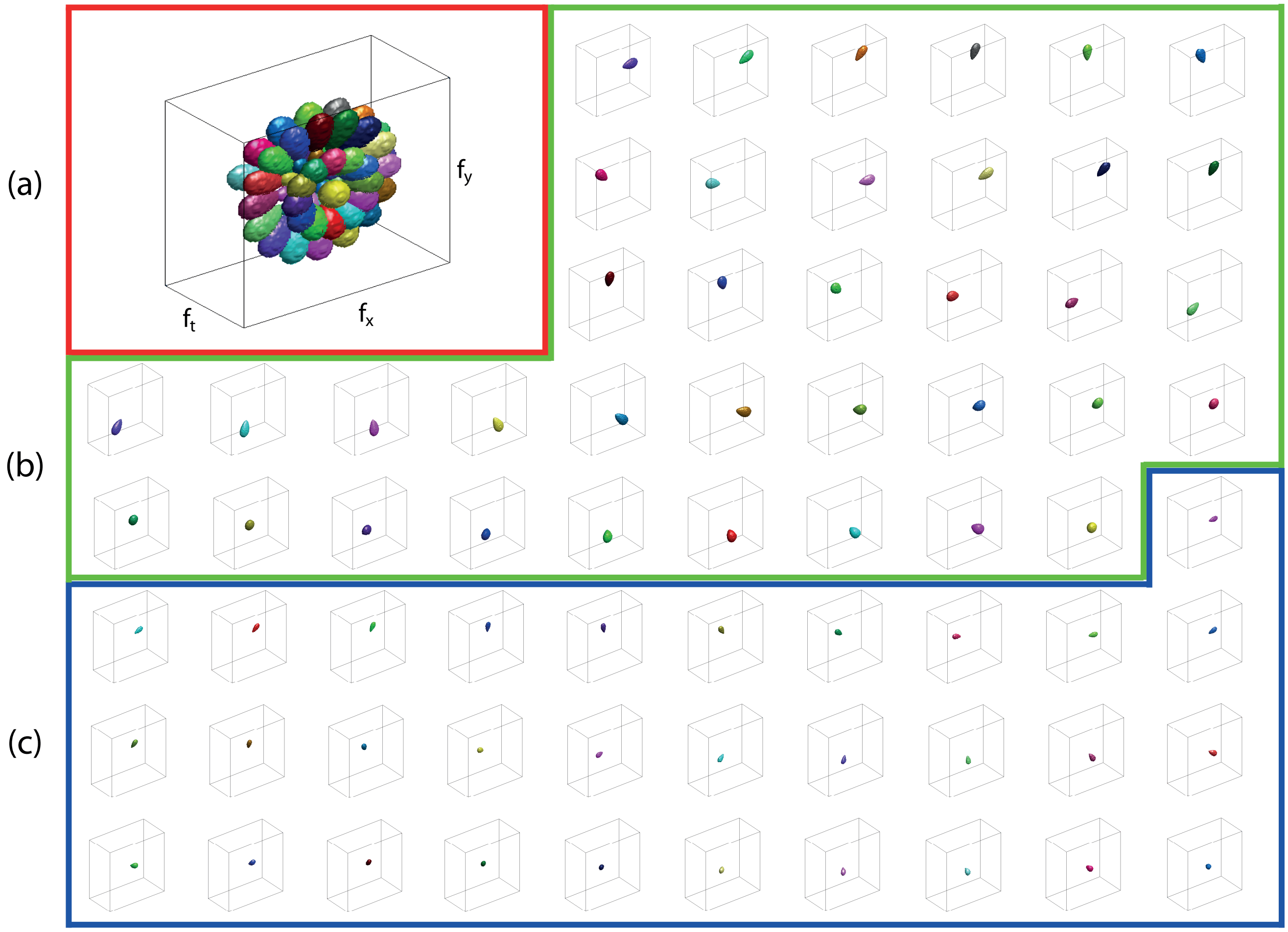


Fig. 5 Visualization of the filters in 3-D: all filters from the first scale **(b)** and the second scale **(c)** are shown together in **a**. (For visualization, we specified a cutoff at 3 dB on the filters)

the motion and scene components which are normal to the orientation of the filter are pronounced, as illustrated in the example in Fig. 4.

3 Implementation

A flowchart describing the implementation of our method is depicted in Fig. 2. Our goal is to represent each video sequence by a single global descriptor and perform the classification. For the current implementation, we extract K uniformly sampled clips of a fixed length from each given video.

As the second step, we compute the 3-D DFT and obtain the frequency spectrum of each clip as given by Eq. 2. In order to capture the components at various intervals of the frequency spectrum of a clip, we apply a bank of narrow band 3-D Gabor filters with different orientations and scales. The transfer function of each 3-D filter, tuned to a spatial frequency f_{r_0} along the direction specified by the polar and the azimuthal orientation angles θ_0 and ϕ_0 in a spherical coordinate system, can be expressed by

$$G(f_r,\theta,\phi) = exp\left\{-\frac{(f_r - f_{r_0})^2}{2\sigma_r^2} - \frac{(\theta-\theta_0)^2}{2\sigma_\theta^2} - \frac{(\phi-\phi_0)^2}{2\sigma_\phi^2}\right\}, \tag{6}$$

where $f_r = \sqrt{f_x^2 + f_y^2 + f_t^2}$, $\theta = arctan\left(\frac{f_y}{f_x}\right)$ and $\phi = arccos\left(\frac{f_z}{\sqrt{f_x^2+f_y^2+f_t^2}}\right)$. The parameters σ_r, σ_θ and σ_ϕ are the radial and angular bandwidths, respectively, defining the elongation of the filter in the spatio-temporal frequency domain.

The combination of our filters is selected to cover only half of the total volume of the frequency spectrum due to the symmetrical nature of the discrete Fourier transform. 3-D plots of these filters are shown in Fig. 5. Applying each generated 3-D filter on the frequency spectrum of the clip, we compute the output

$$\Gamma_i(f_x,f_y,f_t) = F(f_x,f_y,f_t)\ [G_i(f_x,f_y,f_t)], \tag{7}$$

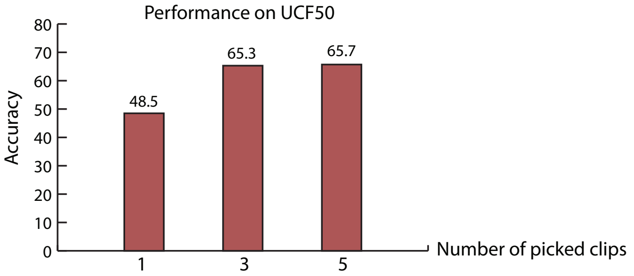


Fig. 6 Effect of number of picked clips on the classification performance

where $\mathit{\Gamma}_i(f_x, f_y, f_t)$ is the output when the i^{th} filter is applied. Then we take the inverse 3-D DFT

$$H_i(x,y,t)=\sum_{f_x=0}^{M-1}\sum_{f_y=0}^{N-1}\sum_{f_t=0}^{T-1}\mathit{\Gamma}_i(f_x,f_y,f_t)e^{j2\Pi(\frac{xf_x}{M}+\frac{yf_y}{N}+\frac{tf_t}{T})}. \qquad (8)$$

By quantizing the output volume in fixed sub-volumes and taking the sum of each sub-volume and performing the same computation for each filter in our filter bank, we obtain a long feature vector which represents a single clip. This feature vector has the advantage of preserving the spatial information as the response of each filter on each sub-volume contributes to an element in the concatenated feature vector. The last step is to apply PCA, a popular method for dimensionality reduction, in order to generate our global video descriptor.

4 Experimental Results

To test the performance of our approach, we used publicly available datasets: KTH, UCF50, and HMDB51. UCF50 and HMDB51 are the two most challenging datasets with the largest number of classes, which are collections of thousands of low-quality web videos with camera motion, different viewing directions, large interclass variations, cluttered backgrounds, occlusions and varying illumination conditions. For all experiments, we picked three key clips of 64 frames from each video and downsampled the frames of clips to a fixed size (128 × 128) for computational efficiency. Picking more than three clips did not result in a further improvement in performance as depicted in Fig.6; hence, we report results based on three clips. Next, we computed the 3-D DFT to compute the frequency spectrum of the clips of each video and then applied the generated filter bank.

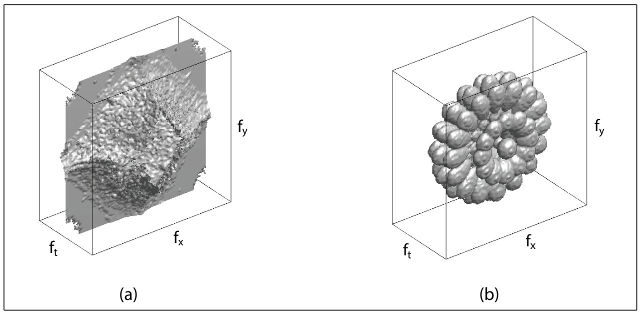


Fig. 7 The cumulative power spectrum of 500 sample videos **(a)** is captured effectively by the selection of our 3-D filter bank **(b)**

Our generated filter bank, described in Sect. 3, consisted of 68 3-D Gabor filters, which corresponded to 2 scales and 37 and 31 orientations for the first and second scales, respectively, in the spatio-temporal frequency domain. The filters are shown in Fig. 5. The selection of filters was designed experimentally to capture the frequency components effectively as shown in Fig. 7; the cumulative power spectrum of 500 videos of different actions was computed and our filter bank captures more than 99 % of the total power. There was no need for another set of filters, which captures very high frequencies with negligible power. However, increasing the number of filters in the pass band makes the filters become narrower and the descriptor to have a finer response, with a penalty of higher computational requirements. As an experiment, we tested a three-scale filter set with 64 narrower filters per scale and obtained an additional 2.5 % performance improvement on UCF50. Considering the computation time trade-off, we did not use this configuration for the reported results.

In our experiments, the central frequencies for the two scales of filters were set to 38.8 and 19 with radial bandwidths 14.2 and 8.6, respectively. The angular bandwidths σ_θ and σ_ϕ were set to 0.2 and 0.1, respectively. Each of the filters we computed had $128\times128\times64$ as the size of the frequency spectrum of clips. After the application of the filters, we computed the average response of filters on 512 uniformly spaced $16\times16\times8$ sub-volumes, to quantize and generate the global feature vector for the clip. The length of the feature vector in our experiments was 104,448, as there are 68 filters, 512 sub-volumes and 3 key clips. We reduced the dimensionality of the feature vectors to 2,000 using PCA [23]. To achieve even higher performance, we tested combining our descriptor with a state-of-the-art local descriptor [24], Space−Time Interest Points (STIP). We computed dense STIP features, and generated 1,000 and 2,000-dimensional codebooks to represent each sequence as a histogram.

For classification, we trained a multi-class support vector machine (SVM) [25] using the linear kernel for our descriptor and histogram intersection kernel for STIP.

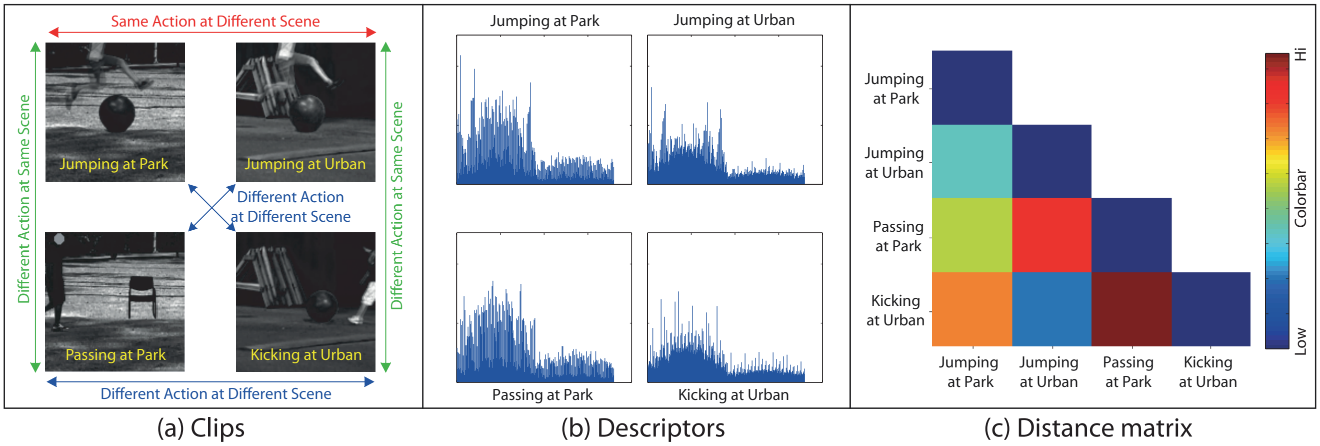


Fig. 8 Descriptor distances for example clips: for the clips with the similarities and differences mentioned in **a**, the distances of the computed descriptors **(b)** are shown as a color-coded matrix in **c**. The descriptors with similar actions and scene have lower distances

We performed cross validation by leaving one group out for testing and training the classifier on the rest of the dataset and performing the same experiment for all groups on UCF50. For HMDB51, we performed cross validation on the three splits of the dataset. We did not include any clips of a video in the test set if any other clip of the same video was used in the training set.

The discriminative power of our descriptor can be seen clearly in the example in Fig. 8. This basic experiment was done using four sequences from a public dataset. For each of the four sequences, we computed the descriptors. Each entry in the matrix in Fig. 8c is the normalized Euclidean distance between the computed descriptors of the four sequences. As seen in the matrix, the descriptor distances between the jumping actions in two different scenes is comparably lower than the other distances, which shows that our descriptor can generalize over intra-class variations. The distances are high when different actions are performed in different scenes, such as the ones labeled by blue arrows in Fig. 8a.

To illustrate the advantage of the 3-D global descriptor, we compare our descriptor with the popular descriptors: GIST [20] (on UCF50 dataset) and STIP[14, 24] (on KTH, UCF50 and HMDB51 datasets) which involve the computation of histograms of oriented gradients (HOG) and histograms of optical flow (HOF). For comparison, we also list the performance of a low-level descriptor based on color and gray values[3] (on UCF50 and HMDB51 datasets), and the biologically motivated C2 features [26,3](on KTH and HMDB51 datasets). Fig. 9 shows the comparison of performance over three datasets.

Table 1 Classification Results of 6 action classes of the KTH dataset

Descriptor	Accuracy
C2 [26]	91.7
STIP [14]	91.8
Gilbert [27]	94.5
GIST3D	**92.0**

4.1 KTH Dataset:

The KTH dataset includes videos captured in a controlled setting of six action classes with 25 subjects for each class. As depicted in Table 1, our descriptor has a classification accuracy of 92.0 %, which is comparable to the state of the art. Fig. 13 shows the confusion table for this dataset. This experiment shows that our descriptor is able to discriminate between the actions with different motions appearing in similar scenes.

4.2 UCF50 Dataset:

This dataset includes unconstrained web videos of 50 action classes with more than 100 videos for each class. As depicted in Table 2, our descriptor has an accuracy of 65.3 % over 50 action classes, which outperforms GIST and STIP. For evaluating the performance of the GIST descriptor, we have used various numbers (3, 20, 40) of sampled frames for each video and performed classification after concatenating the computed descriptors for each frame. The accuracy increased up to 42.4 % when 40 frames were used. Fig. 11 shows the confusion tables for GIST, STIP and our descriptor, GIST3D. Using the combination of STIP and GIST3D by late fusion resulted in a classification accuracy of

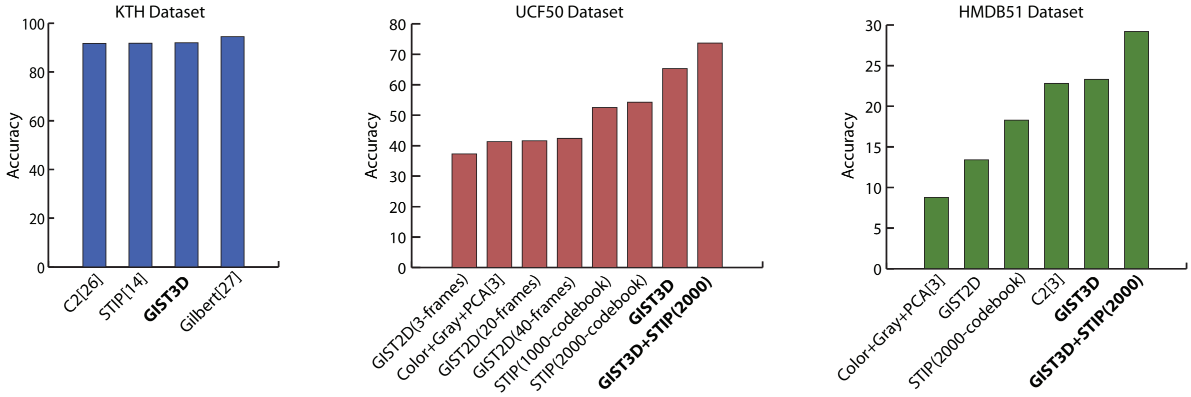


Fig. 9 Average classification accuracies over KTH, UCF50 and HMDB51 datasets

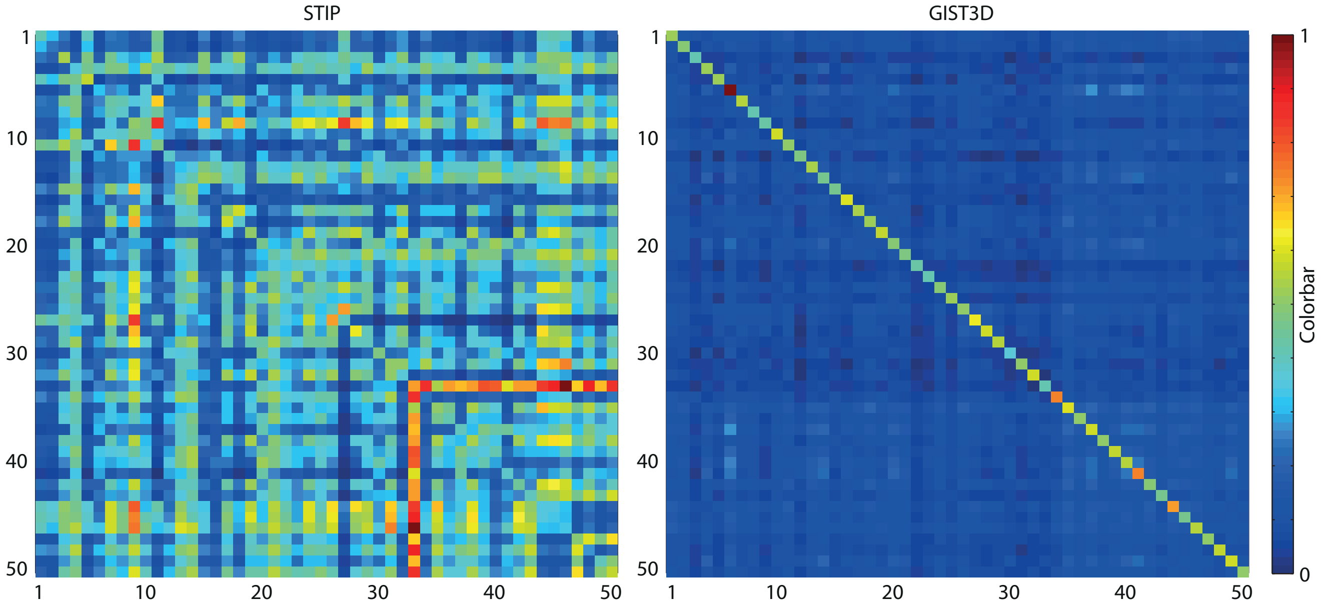


Fig. 10 Descriptor similarity matrices for STIP **(a)** and GIST3D **(b)** computed among 50 action classes of UCF50 dataset

Table 2 Classification Results of 50 action classes of the UCF50 dataset

Descriptor	Accuracy
GIST(3-frames)	37.3
Color+Gray+PCA[3]	41.3
GIST(20-frames)	41.6
GIST(40-frames)	42.4
STIP(HOG/HOF)(1000-dim codebook)	52.5
STIP(HOG/HOF)(2000-dim codebook)	54.3
GIST3D	**65.3**
GIST3D+STIP(2000-dim codebook)	**73.7**

73.7 %, which is another 8 % improvement in the performance. Fig. 11d depicts the confusion table for the combined classification.

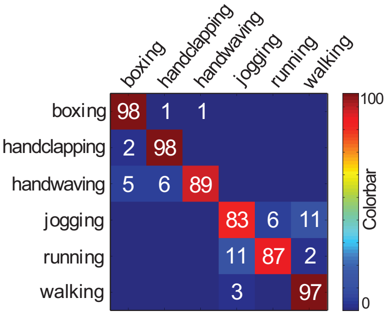
For comparison of our descriptor to STIP, we also analyzed the average similarities of descriptors among action classes of UCF50. We computed the Euclidean similarity for our descriptors and histogram intersection as the similarity measure for STIP. Our descriptor has higher intra-class similarity and lower inter-class similarity than STIP as shown in Fig. 10. This clearly explains why our global descriptor (GIST3D) performs superior than STIP.

4.3 HMDB51 Dataset:

The HMDB51 dataset includes videos of 51 action classes with more than 101 videos for each class. As depicted in Table 3, our descriptor has a classification accuracy of 23.3 % over 51 action classes, which outperforms STIP by 5 %. The late fusion classifier of these two descriptors resulted in a 6 % improvement in the performance over using just our descriptor GIST3D. Fig. 12 shows the confusion tables for STIP, our descriptor and

Table 3 Classification Results of 51 action classes of the HMDB51 dataset

Descriptor	Accuracy
Color+Gray+PCA[3]	8.8
GIST[3]	13.4
STIP(HOG/HOF)(2000-dim codebook)[24]	18.3
C2(Motion+Shape)[3]	22.8
GIST3D	**23.3**
GIST3D+STIP(2000-dim codebook)	**29.2**



	boxing	handclapping	handwaving	jogging	running	walking
boxing	98	1	1			
handclapping	2	98				
handwaving	5	6	89			
jogging				83	6	11
running				11	87	2
walking				3		97

Fig. 13 Confusion Table for KTH using our descriptor, GIST3D

the fused classifier. The actions in the video sequences of HMDB51 are not isolated; multiple actions may be present in a single video sequence despite a given single class label for the sequence. There is also large intra-class scene variation. Therefore, classifying actions on this dataset is more challenging and the performances of the mentioned methods are lower.

4.4 Discussion:

By comparing our classification accuracy with the tested descriptors and analyzing the Tables 2 and 3, we found out that GIST, being a scene descriptor, suffered from the lack of captured motion information. For example, as observed in Fig 11b, GIST cannot differentiate between the actions of Playing Guitar and Playing Violin which happen in similar indoor scenes. The motion is discriminative for these videos and our descriptor (GIST3D) is able to differentiate these actions, as shown in Fig. 11c. Conversely, STIP suffered from locality, as it did not capture the global scene structure and did not carry spatio-temporal information due to the global histogram representation. For example, walking with dog and horse riding actions have similar translating motion, and STIP is not as discriminative as our descriptor for these two actions, as depicted in Fig. 11a. The horse riding videos mostly have rural scenes with periodic vertical and horizontal components such as fences, whereas the walking with dog videos contain urban or park scenes. Our descriptor encodes the useful scene information and is able to discriminate between these two actions.

5 Conclusion

In this paper, we presented a global scene and motion descriptor to classify realistic videos of different actions. Without interest point detection, background subtraction and tracking, we represented each video with a single feature vector and obtained promising classification accuracies using a SVM. Preserving also the useful spatial information, our descriptor had a better performance than the state-of-the-art local descriptor, STIP, utilizing a bag-of-features representation which discards the spatial distribution of the local descriptors. In addition, by combining our descriptor with STIP, we achieved the highest classification accuracies on the challenging datasets, UCF50 and HMDB51. Our descriptor performed comparable to the state-of-the-art descriptors on the KTH dataset, as there was no useful scene information in this dataset. The experiments showed that both scene structure and motion information are important in classifying realistic videos.

Acknowledgements The research presented in this paper is supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center, contract number D11PC20071. The US government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC or the U.S. Government.

References

1. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04). Volume 3. (2004) 32–36
2. (http://vision.eecs.ucf.edu/datasetsActions.html)
3. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: ICCV. (2011)
4. Poppe, R.: A survey on vision-based human action recognition. Image Vision Comput. **28** (2010) 976–990
5. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. Comput. Vis. Image Underst. **115** (2011) 224–241
6. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'08). (2008)

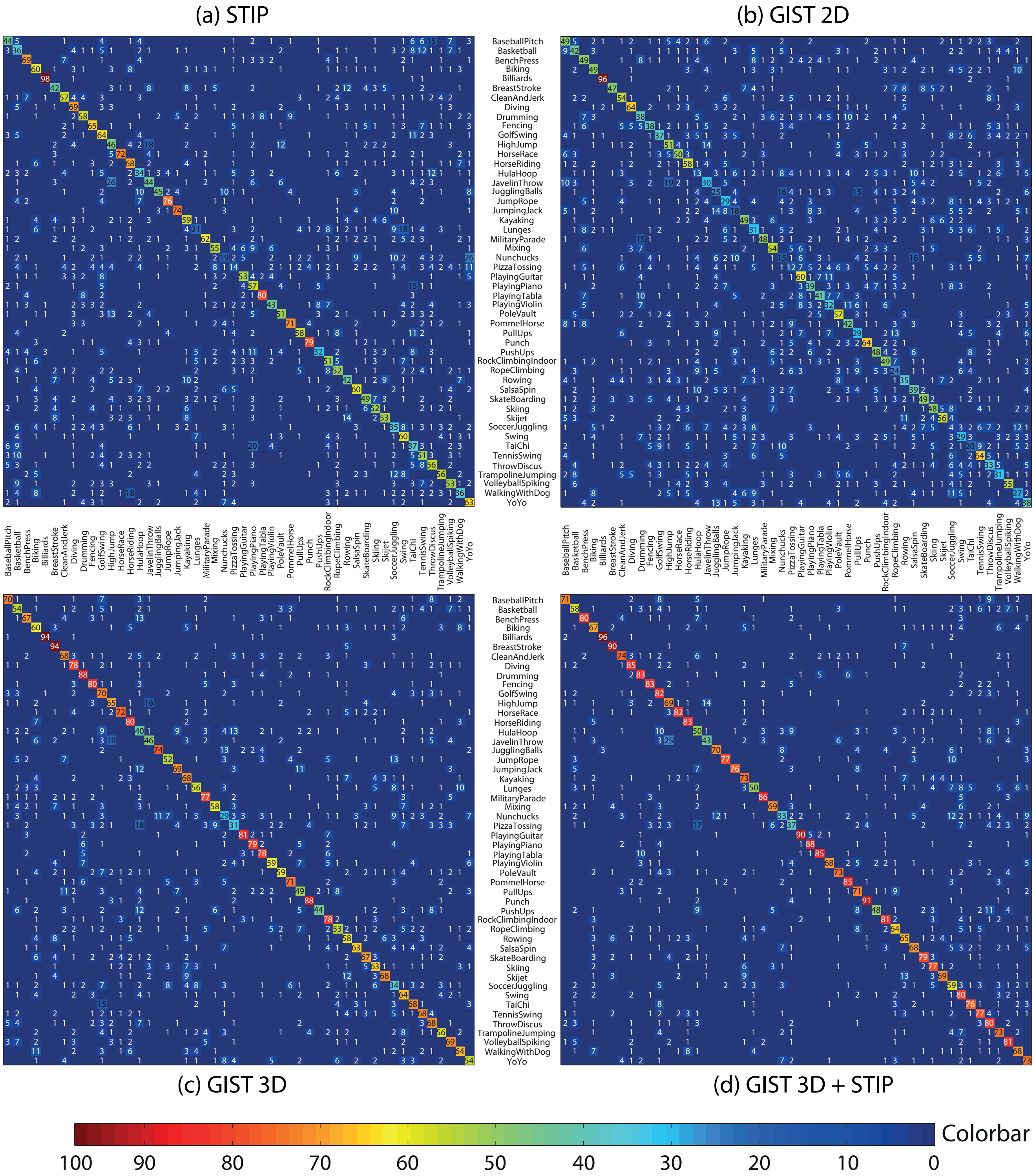


Fig. 11 Confusion tables for STIP **(a)**, GIST(40 images per video)**(b)** and GIST3D **(c)** separately over 50 action classes of UCF50 dataset (Average accuracies for GIST, STIP and GIST3D are 42.4, 54.3 and 65.3 %, respectively.) Combining GIST3D and STIP resulted in the confusion table in **d** with an average accuracy of 73.7 %)

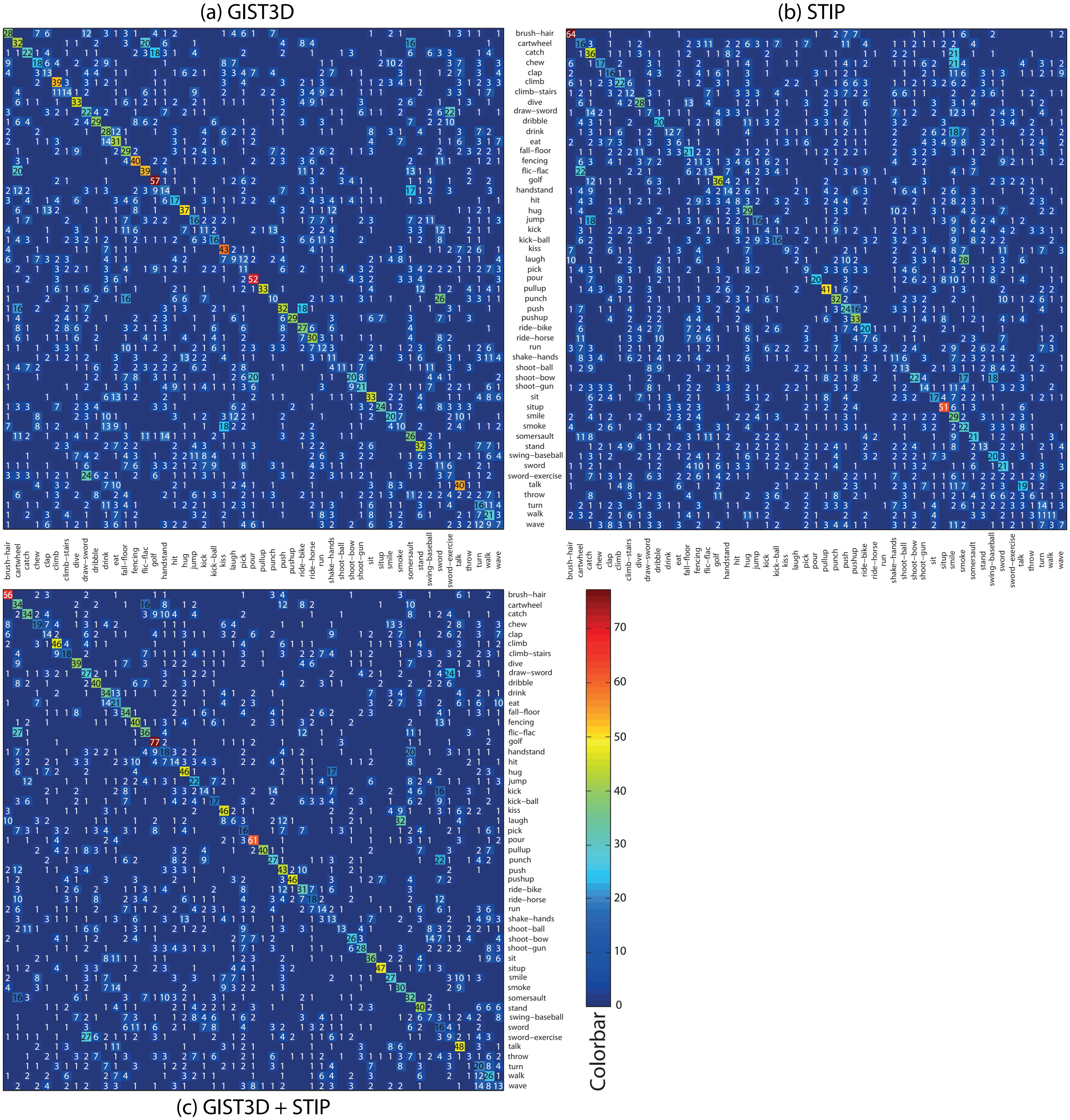


Fig. 12 Confusion tables for GIST3D **(a)**, and STIP **(b)**, and the combined classifier **(c)** over HMDB51 dataset. (Average accuracies for STIP, GIST3D, and the combined classifier GIST3D+STIP are 18.3, 23.3, and 29.2 %, respectively)

7. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09). (2009) 1996 –2003
8. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (2001) 257–267
9. Yilmaz, A., Shah, M.: A differential geometric approach to representing the human actions. Comput. Vis. Image Underst. **109** (2008) 335–351
10. Black, M.: Explaining optical flow events with parameterized spatio-temporal models. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'99). Volume 1. (1999) 326–332
11. Polana, R., Nelson, R.C.: Detection and recognition of periodic, non-rigid motion. International Journal of Computer Vision **23** (1997) 261–282
12. Wu, S., Oreifej, O., Shah, M.: Action recognition in videos acquired by a moving camera using motion decom-

position of lagrangian particle trajectories. In: IEEE International Conference on Computer Vision (ICCV'11). (2011) 1419 –1426

13. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11). (2011) 3169 –3176
14. Laptev, I.: On space-time interest points. Int. J. Comput. Vision **64** (2005) 107–123
15. Harris, C., Stephens, M.: A combined corner and edge detector. In: In Proc. of Fourth Alvey Vision Conference. (1988) 147–151
16. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS. (2005) 65–72
17. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC. (2008)
18. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: ACM Multimedia. (2007) 357–360
19. Ikizler-Cinbis, N., Sclaroff, S.: Object, scene and actions: combining multiple features for human action recognition. In: Proceedings of the 11th European conference on Computer vision (ECCV'10). (2010) 494–507
20. Oliva, A., Torralba, A.B., Guerin-Dugue, A., Herault, J.: Global semantic classification of scenes using power spectrum templates. Challenge of Image Retrieval (1999) 1–12
21. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vision **42** (2001) 145–175
22. Heeger, D.J.: Notes on motion estimation, http://white.stanford.edu/∼heeger, (1998)
23. Maaten, L.V.D., Postma, E.O., Herik, H.J.V.D.: Dimensionality reduction: A comparative review (2008)
24. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC. (2009) 127
25. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27
26. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: IEEE 11th International Conference on Computer Vision (ICCV'07). (2007) 1–8
27. Gilbert, A., Illingworth, J., Bowden, R.: Action recognition using mined hierarchical compound features. IEEE Transactions on Pattern Analysis and Machine Intelligence **33** (2011) 883–897