Machine Vision and Applications (2013) 24:971–981
DOI 10.1007/s00138-012-0450-4

ORIGINAL PAPER

Recognizing 50 human action categories of web videos

Kishore K. Reddy · Mubarak Shah

Received: 22 January 2012 / Revised: 5 September 2012 / Accepted: 7 September 2012 / Published online: 16 November 2012

Abstract Action recognition on large categories of unconstrained videos taken from the web is a very challenging problem compared to datasets like KTH (6 actions), IXMAS (13 actions), and Weizmann (10 actions). Challenges like camera motion, different viewpoints, large interclass variations, cluttered background, occlusions, bad illumination conditions, and poor quality of web videos cause the majority of the state-of-the-art action recognition approaches to fail. Also, an increased number of categories and the inclusion of actions with high confusion add to the challenges. In this paper, we propose using the scene context information obtained from moving and stationary pixels in the key frames, in conjunction with motion features, to solve the action recognition problem on a large (50 actions) dataset with videos from the web. We perform a combination of early and late fusion on multiple features to handle the very large number of categories. We demonstrate that scene context is a very important feature to perform action recognition on very large datasets. The proposed method does not require any kind of video stabilization, person detection, or tracking and pruning of features. Our approach gives good performance on a large number of action categories; it has been tested on the UCF50 dataset with 50 action categories, which is an extension of the UCF YouTube Action (UCF11) dataset containing 11 action categories. We also tested our approach on the KTH and HMDB51 datasets for comparison.

Keywords Action recognition · Web videos · Fusion

K. K. Reddy (✉) · M. Shah
4000 Central Florida Blvd, Orlando, USA
e-mail: kkreddy@mail.ucf.edu

M. Shah
e-mail: shah@crcv.ucf.edu

1 Introduction

Action recognition has been a widely researched topic in computer vision for over a couple of decades. Its applications in real-time surveillance and security make it more challenging and interesting. Various approaches have been taken to solve the problem of action recognition [20]; however, the majority of the current approaches fail to address the issue of a large number of action categories and highly unconstrained videos taken from web.

Most state-of-the-art methods developed for action recognition are tested on datasets like KTH, IXMAS, and Hollywood (HOHA), which are largely limited to a few action categories and typically taken in constrained settings. The KTH and IXMAS datasets are unrealistic; they are staged, have minor camera motion, and are limited to less than 13 actions which are very distinct. The Hollywood dataset [9], which is taken from movies, addresses the issue of unconstrained videos to some extent, but involves actors, contains some camera motion and clutter, and is shot by a professional camera crew under good lighting conditions. The UCF YouTube Action (UCF11) dataset [10] consists of unconstrained videos taken from the web and is a very challenging dataset, but it has only 11 action categories, all of which are very distinct actions. The UCF50 dataset, which is an extension of the UCF11 dataset, also contains videos downloaded from YouTube and has 50 action categories. The recently released HMDB51 dataset [8] has 51 action categories, but after excluding facial actions like smile, laugh, chew, and talk, which are not articulated actions, it has 47 categories compared to 50 categories in UCF50. Most of the current methods would fail to detect an action/activity in datasets like UCF50 and HMDB51 where the videos are taken from web. These videos contain random camera motion, poor lighting conditions, clutter, as well as changes in scale, appearance,

Table 1 Action datasets

Datasets	Number of actions	Camera motion	Background
KTH	6	Slight motion	Static
Weizmann	10	Not present	Static
IXMAS	14	Not present	Static
UCF sports	9	Present	Dynamic
HOHA	12	Present	Dynamic
UCF11	11	Present	Dynamic
UCF50	50	Present	Dynamic
HMDB51	51 (47)	Present	Dynamic

and viewpoints, and occasionally no focus on the action of interest. Table 1 shows the list of action datasets.

In this paper, we study the effect of large datasets on performance, and propose a framework that can address issues with real-life action recognition datasets (UCF50). The main contributions of this paper are as follows:

1. We provide an insight into the challenges of large and complex datasets like UCF50.
2. We propose the use of moving and stationary pixel information obtained from optical flow to obtain our scene context descriptor.
3. We show that as the number of actions to be categorized increases, the scene context plays a more important role in action classification.
4. We propose the idea of early fusion schema for descriptors obtained from moving and stationary pixels to understand the scene context, and finally perform a probabilistic fusion of scene context descriptor and motion descriptor.

To the best of our knowledge, no one has attempted action/activity recognition on such a large-scale dataset (50 action categories) consisting of videos taken from the web (unconstrained videos) using only visual information.

The rest of the paper is organized as follows. Section 2 deals with the related work. Section 3 gives an insight into working with large datasets. In Sects. 4 and 5, we introduce our proposed scene context descriptor and the fusion approach. In Sect. 6, we present the proposed approach, followed by the experiments and results with discussions in Sect. 7. Finally, we conclude our work in Sect. 8.

2 Related work

Over the past two decades, a wide variety of approaches has been tried to solve the problem of action recognition. Template-based methods [1], modeling the dynamics of human motion using finite state models [6] or hidden Markov models [21], and Bag of Features models [4,10,11,22] (BOF) are a few well-known approaches taken to solve action recognition. Most of the recent work has been focused on BOF in one form or another. However, most of this work is limited to small and constrained datasets.

Categorizing large numbers of classes has always been a bottleneck for many approaches in image classification/action recognition. Deng et al. [3] demonstrated the challenges of doing image classification on 10,000 categories. Recently, Song et al. [17] and Zhao et al. [19] attempted to categorize large numbers of video categories by using text, speech, and static and motion features. Song et al. [17] used visual features like color histogram, edge features, face features, SIFT, and motion features and showed that text and audio features outperform visual features by a significant margin.

With the increase in number of action categories, motion features alone are not discriminative enough for reliable action recognition. Marszalek et al. [13] introduced the concept of context in action recognition by modeling the scenes. 2D-Harris detector is used to detect salient regions from which SIFT descriptor is extracted and bag-of-features framework is used to obtain the static appearance descriptor. Han et al. [5] detects person, body parts, and the objects involved in an action and used the knowledge of their spatial location to design contextual scene descriptor. Recently, Choi et al. [2] introduced the concept of "Crowd Context" to classify activities involving interaction between multiple people. In all the proposed methods [2,5,13], the performance depends on the detectors used.

Extracting reliable features from unconstrained web videos has been a challenge. In recent years, action recognition in realistic videos was addressed by Laptev et al. [9] and Liu et al. [10,11]. Liu et al. [10] proposed pruning of the static features using PageRank and motion features using motion statistics. Fusion of these pruned features showed a significant increase in the performance on the UCF11 dataset. Ikizler et al. [7] used multiple features from the scene, object, and person, and combined them using a Multiple MIL (multiple instance learning) approach. Fusion of multiple features extracted from the same video has gained significant interest in recent years. Work by Snoek et al. [14] compares early and late fusion of descriptors.

There has been no action recognition work done on very large datasets, using only visual features. In this paper, we propose a method which can handle these challenges.

3 Analysis on large-scale dataset

UCF50 is the largest action recognition dataset publicly available, after excluding the non-articulated actions from the HMDB51 dataset. UCF50 has 50 action categories with a total of 6676 videos, and with a minimum of 100 videos for

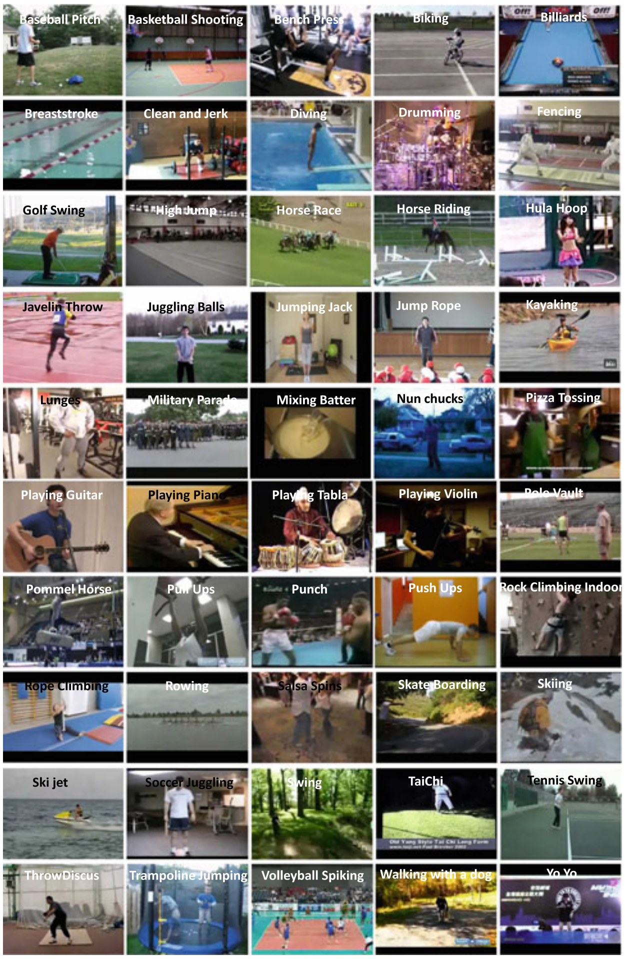


Fig. 1 *Screenshots* from videos in the UCF50 dataset showing the diverse action categories

each action class. Samples of video screenshots from UCF50 are shown in Fig. 1. This dataset is an extension of UCF11. In this section, we perform a baseline experiment on UCF50 by extracting the motion descriptor and using the bag of video words approach. We use two classification approaches:

1. **BoVW-SVM:** support vector machines (SVM) to do classification.
2. **BoVW-NN:** nearest neighbor approach using SR-Tree to do classification.

Which motion descriptor do we use?

Due to the large scale of the dataset, we prefer a motion descriptor which is faster to compute and reasonably accurate. To decide on the motion descriptor, we performed experiments on a smaller dataset KTH with different motion descriptors, which were extracted from the interest points detected using Dollar's detector [4]. At every interest point location (x, y, t) , we extract the following motion descriptors:

Table 2 Performance of different motion descriptors on the KTH dataset

Method	Codebook (%)		
	100	200	500
Gradient	83.78	87.62	90.13
Optical flow	85.64	87.12	90.15
3D-SIFT	85.11	88.65	91.13

- **Gradient:** At any given interest point location in a video (x, y, t), a 3D cuboid is extracted. The brightness gradient is computed in this 3D cuboid, which gives rise to three channels (G_x, G_y, G_t) that are flattened into a vector, and later PCA is applied to reduce the dimension.
- **Optical flow:** Similarly, Lucas–Kanade optical flow [12] is computed between consecutive frames in the 3D cuboid at (x, y, t) location to obtain two channels (V_x, V_y). The two channels are flattened and PCA is utilized to reduce the dimension.
- **3D-SIFT:** Three-dimensional SIFT proposed by Scovanner et al. [15] is an extension of SIFT descriptor to spatio-temporal data. We extract 3D-SIFT around the spatio-temporal region of a given interest point (x, y, t).

All of the above descriptors are extracted from the same location of the video and the experimental setup is identical. We use BOF paradigm and SVM to evaluate the performance of each descriptor. From Table 2, one can notice that 3D-SIFT outperforms the other two descriptors for codebook of size 500, whereas gradient and optical flow descriptors perform the same. Computationally, the gradient descriptor is the fastest and 3D-SIFT is the slowest. Due to the time factor, we will use gradient descriptor as our motion descriptor for all further experiments.

We also tested our framework on the recently proposed motion descriptor MBH by Wang et al. [18]. The MBH descriptor encodes the motion boundaries along the trajectories obtained by tracking densely sampled points using optical flow fields. Using the code provided by the authors [18], MBH descriptors are extracted for UCF11 and UCF50 datasets and used in place of the above-mentioned motion descriptor for comparison of results with [18].

3.1 Effect of increasing the action classes

In this experiment, we show that increasing the number of action classes affects the recognition accuracy of a particular action class. Since the UCF11 dataset is a subset of UCF50, we first start with the 11 actions from the UCF11 dataset and randomly add new actions from the remaining 39 different actions from the UCF50 dataset. Each time a new

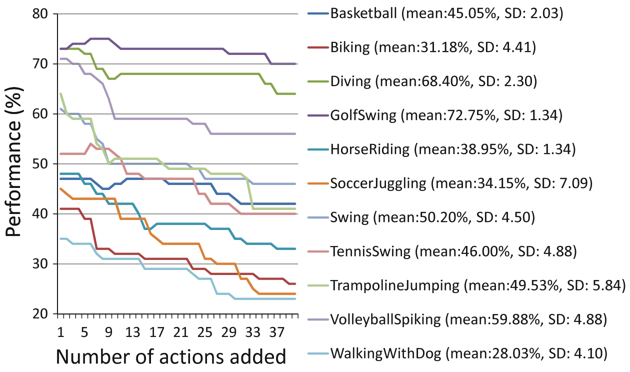


Fig. 2 The effect of increasing the number of actions on the UCF YouTube Action dataset's 11 actions by adding new actions from UCF50 using only the motion descriptor. Standard deviation (SD) and mean are also shown next to the action name. The performance on the initial 11 actions decreases as new actions are added

action is added, a complete leave-one-out cross validation is performed using bag of video words approach on motion descriptor and SVM for classification on the incremented dataset using a 500-dimension codebook. Performance using BoVW-SVM on the initial 11 actions is 55.46 % and BoVW-NN is 37.09 %. Even with the increase in the number of actions in the dataset, SVM performs significantly better than the nearest neighbor approach.

Figure 2 shows the change in performance by using BoVW-SVM on the initial 11 actions as we add the 39 new actions, one at a time. Increasing the number of actions in the dataset has affected some actions more than others. Actions like "soccer juggling" and "trampoline jumping" were most affected; they have a standard deviation of ∼7.08 and ∼5.84 %, respectively. Some actions like "golf swing" and "basketball" were least affected with a very small standard deviation of ∼1.35 and ∼2.03 %, respectively. Overall, the performance on 11 actions from UCF11 dropped by ∼13.18 %, i.e., from 55.45 to 42.27 %, by adding 39 new actions from UCF50. From Fig. 2, one can also notice that 8 of 11 actions have standard deviation of more than ∼4.10 %. Analysis of the confusion table shows a significant confusion of these initial 11 actions with newly added actions. This shows that the motion feature alone is not discriminative enough to handle more action categories.

To address the above concerns, we propose a new scene context descriptor which is more discriminative and performs well in very large action datasets with a high number of action categories. From the experiments on UCF50, we show that the confusion between actions is drastically reduced and the performance of the individual categories increased by fusing the proposed scene context descriptor.

4 Scene context descriptor

In order to overcome the challenges of unconstrained web videos and handle a large dataset with lots of confusing actions, we propose using the scene context information in which the action is happening. For example, skiing and skateboarding, horse riding and biking, and indoor rock climbing and rope climbing have similar motion patterns with high confusion, but these actions take place in different scenes and contexts. Skiing happens on snow, which is very different from where skateboarding is done. Similarly, horse riding and biking happen in very different locations. Furthermore, scene context also plays an important role in increasing the performance on individual actions. Actions are generally associated with places, e.g., diving and breast stroke occur in water, and golf and javelin throw are outdoor sports. In order to increase the classification rate of a single action, or to reduce the confusion between similar actions, the scene information is crucial, along with the motion information. We refer to these places or locations as scene context in our paper.

As the number of categories increases, the scene context becomes important, as it helps reduce the confusion with other actions having similar kinds of motion. In our work, we define scene context as the place where a particular motion happens (stationary pixels), and also include the object that creates this motion (moving pixels).

Humans have an extraordinary ability to perform object detection, tracking and recognition. We assume that humans tend to focus on objects that are salient or the things that move in their field of view. We try to mimic this by coming up with groups of moving pixels which can be roughly assumed as salient regions and groups of stationary pixels as an approximation of non-salient regions in a given video.

Moving and stationary pixels: Optical flow gives a rough estimate of velocity at each pixel given two consecutive frames. We use optical flow (u, v) at each pixel obtained using Lucas–Kanade method [12] and apply a threshold on the magnitude of the optical flow to decide if the pixel is moving or stationary. Figure 3 shows the moving and stationary pixels in several sample key frames. We extract dense CSIFT [14] at pixels from both groups and use BOF paradigm to get a histogram descriptor for both groups separately. We performed experiments using CSIFT descriptor, extracted on a dense sampling of moving pixels MP_v and stationary pixels SP_v. For a 200-dimension codebook, the moving pixels CSIFT histogram alone resulted in a 56.63 % performance, while the stationary pixels CSIFT histogram achieved 56.47 % performance on the UCF11. If we ignore the moving and stationary pixels and consider the whole image as one, we obtain a performance of 55.06 %. Our experiments show that concatenation of histogram descriptors of moving and stationary pixels using CSIFT gives the best performance of 60.06 %. From our results, we conclude that concatenation of MP_v and SP_v into one descriptor SC_v is a very unique way to encode the scene context information. For example, in a diving video, the moving pixels are

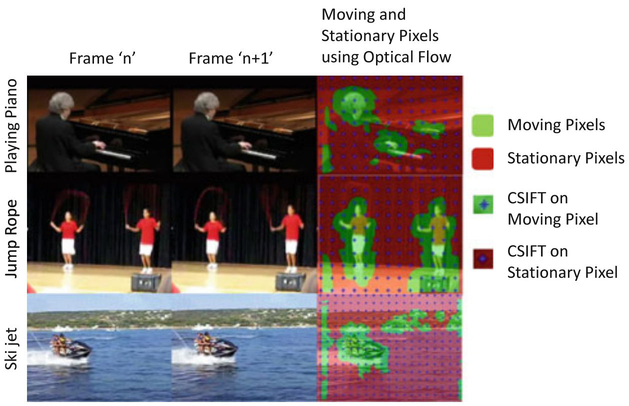


Fig. 3 Moving and stationary pixels obtained using optical flow

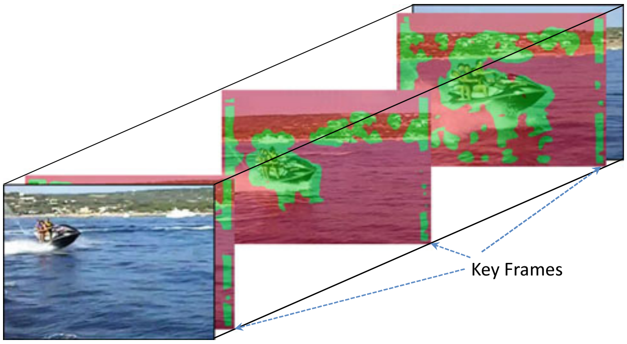


Fig. 4 Key frame selection from a given video

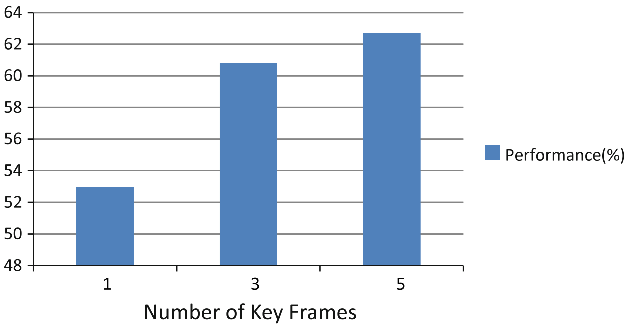


Fig. 5 Performance of scene context descriptor on different number of key frames

mostly from the person diving, and the stationary pixels are mostly from the water (pool), which implies that diving will occur only in water and that this unique scene context will help detect the action diving.

Why CSIFT? Liu et al. [10] show that using SIFT on the UCF11 dataset gave them 58.1 % performance. Our experiments on the same dataset using GIST gave us a very low performance of 43.89 %. Our approach of scene context descriptor using CSIFT gave us a performance of 63.75, ∼2.5 % better than motion feature and ∼5.6 % better than SIFT. It is evident that color information is very important for capturing the scene context information.

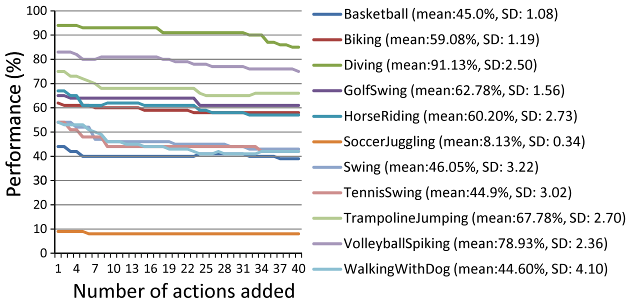


Fig. 6 Effect of increasing the number of actions on the UCF YouTube Action dataset's 11 actions by adding new actions from UCF50, using only the scene context descriptor. Standard deviation (SD) and mean are shown next to the action name.The performance on the initial 11 actions decreases as new actions are added, but with significantly less standard deviation compared to using motion descriptor as shown in Fig. 2

Key frames: Instead of computing the moving and stationary pixels and their corresponding descriptor on all the frames in the video, we perform a uniform sampling of k frames from a given video, as shown in Fig. 4. This reduces the time taken to compute the descriptors, as the majority of the frames in the video are redundant. We did not implement any kind of key frame detection, which can be done by computing the color histogram of frames in the video and considering a certain level of change in color histogram as a key frame. We tested on the UCF11 dataset by taking different numbers of key frames sampled evenly along the video. Figure 5 shows that the performance on the dataset is almost stable after three key frames. In our final experiments on the datasets, we consider three key frames equally sampled along the video to speed up the experiments. In this experiment, a codebook of dimension 500 is used.

4.1 How discriminative is the scene context descriptor?

In this experiment, the proposed scene context descriptors are extracted and a bag of video word paradigm followed by SVM classification is employed to study the proposed descriptor. Similar to the experiment in Sect. 3.1, one new action is added to UCF11 incrementally from UCF50, at each increment leave-one-out cross-validation is performed. The average performance on the initial 11 actions of UCF11 is 60.09 %; after adding 39 new actions from UCF50 the performance on the 11 actions dropped to 52.36 %, i.e., a ∼7.72 % decrease in performance, compared to ∼13.18 % decrease for motion descriptor. The average standard deviation of the performance of the initial 11 actions over the entire experimental setup is ∼2.25 % compared to ∼4.18 % for motion descriptor. Figure 6 clearly shows that the scene context descriptor is more stable and discriminative than the motion descriptor with the increase in the number of action categories.

5 Fusion of descriptors

A wide variety of visual features can be extracted from a single video, such as motion features (e.g., 3DSIFT, spatio-temporal features), scene features (e.g., GIST), or color features (e.g., color histogram). To do the classification using all these different features, the information has to be fused eventually. According to Snoek et al. [16], fusion schemes can be classified into early fusion and late fusion based on when the information is combined.

Early fusion: In this scheme, the information is combined before training a classifier. This can be done by concatenating the different types of descriptors and then training a classifier.

Late fusion: In this scheme, classifiers are trained for each type of descriptor, and then the classification results are fused. Classifiers, such as SVM, can provide a probability estimate for all the classes rather than a hard classification decision. The concept of fusing this probability estimate is called probabilistic fusion [23]. For probabilistic fusion, the different descriptors are considered to be conditionally independent. This is a fair assumption for the visual features that we use in this paper, i.e., motion descriptor using gradients and color SIFT. In probabilistic fusion, the individual probabilities are multiplied and normalized. For d sets of descriptors $\left\{X_j\right\}_1^d$ extracted from a video, the probability of the video being classified as action a, i.e., $p(a\,|\{X_j\}_1^d)$, using probabilistic fusion is:

$$p\left(a\,|\{X_j\}_1^d\right) = \frac{1}{N}\prod_{j=1}^{d} p(a\,|X_j), \tag{1}$$

where N is a normalizing factor which we consider to be 1. In late fusion, the individual strengths of the descriptors are retained.

5.1 Probabilistic fusion of motion and scene context descriptor

Probabilistic fusion: Late fusion using probabilistic fusion requires combining the probability estimates of both the descriptors from their separately trained SVMs, i.e.,

$$max\ (P_{\mathrm{SC}}(i)\ P_{\mathrm{M}}(i))\ ,\ \text{where } i = 1\ to\ a,$$

where a is the number of actions to classify, and $P_{\mathrm{SC}}(i)$ and $P_{\mathrm{M}}(i)$ are the probability estimates of action i, obtained using SVMs trained on scene context descriptors and motion descriptors separately. We also tested early fusion of both motion and scene context features, i.e., $[M_v\ \mathrm{SC}_v]$, and trained an SVM, which gave ∼5 % performance better than individual descriptors on UCF50, which was expected. However, performing an early fusion after normalization, i.e., $[M_v/\max(M_v)\ , \mathrm{SC}_v/\max(\mathrm{SC}_v)]$, gave a remarkable increase in the performance, ∼14 %. It is evident from Fig. 7 that on average across all the codebooks, late fusion (probabilistic fusion) is the best. Therefore, in all of our experiments on KTH, HMDB51, UCF YouTube (UCF11), and UCF50 datasets, we do probabilistic fusion of both scene context and motion descriptors.

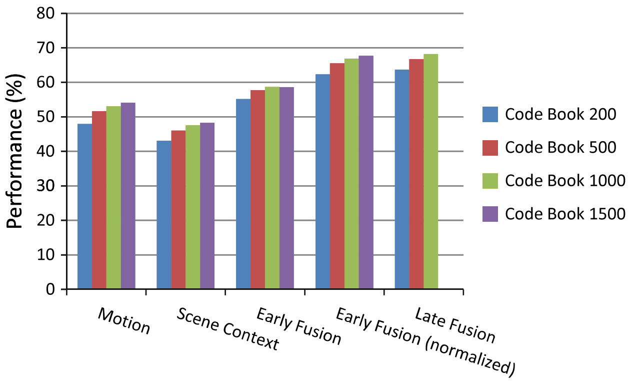


Fig. 7 Performance of different methods to fuse scene context and motion descriptors on the UCF50 dataset

6 System overview

To perform action recognition, we extract the following information from the video: (1) scene context information in key frames and (2) motion features in the entire video, as shown in Fig. 8. The individual SVMs probability estimates are fused to get the final classification.

In the training phase, from each training videos, we extract spatio-temporal features $\{m_1, m_2, \ldots, m_x\}$, from x interest points detected using the interest point detector proposed by Dollar et al. [4]. We also extract CSIFT features on moving pixels $\left\{mp_1, mp_2, \ldots, mp_y\right\}$ and stationary pixels $\{sp_1, sp_2, \ldots, sp_z\}$ from k frames uniformly sampled in the video, where y and z are the number of CSIFT features extracted from moving and stationary regions, respectively. A codebook of size p is generated of the spatio-temporal features from all the training videos. Similarly, a codebook of size q is generated of CSIFT features from moving pixels and stationary pixels combined. For a given video v, we compute the histogram descriptors M_v, MP_v, and SP_v using their respective codebooks for the x spatio-temporal features from the entire video, y CSIFT features from the moving pixels, and z CSIFT features from the stationary pixels from key frames. We do an early fusion of MP_v and SP_v before training a classifier using support vector machine (SVM), i.e., $\mathrm{SC}_v = [\mathrm{MP}_v\ \mathrm{SP}_v]$, which we call the scene context descriptor. We train SVM classifier SVM_M for all the motion descriptors M_v and separate SVM classifier SVM_C for all scene context descriptors SC_v, where $v = [1, 2, \ldots, tr]$ and tr is the number of training videos. Since all the descriptors M_v, MP_v, and SP_v are histograms, we use histogram intersection kernel in the SVM classifier.

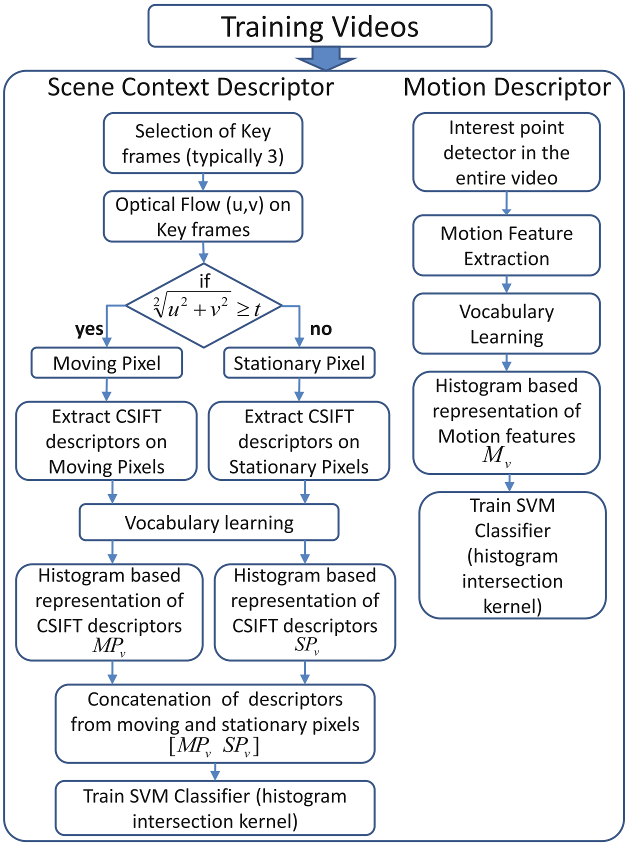

Training Videos
Scene Context Descriptor
Motion Descriptor
Selection of Key frames (typically 3)
Optical Flow (u,v) on Key frames
if $\sqrt[2]{u^2+v^2} \geq t$
yes
no
Moving Pixel
Stationary Pixel
Extract CSIFT descriptors on Moving Pixels
Extract CSIFT descriptors on Stationary Pixels
Vocabulary learning
Histogram based representation of CSIFT descriptors MP_v
Histogram based representation of CSIFT descriptors SP_v
Concatenation of descriptors from moving and stationary pixels $[MP_v \; SP_v]$
Train SVM Classifier (histogram intersection kernel)
Interest point detector in the entire video
Motion Feature Extraction
Vocabulary Learning
Histogram based representation of Motion features M_v
Train SVM Classifier (histogram intersection kernel)

Fig. 8 Proposed approach

Given a query video q, we extract the motion descriptor M_q and the scene context descriptor SC_q, as described in the training phase. We perform a probabilistic fusion of the probability estimates of the motion descriptor $[P_\mathrm{M}(1), P_\mathrm{M}(2), \ldots, P_\mathrm{M}(a)]$, and scene context descriptor $[P_\mathrm{SC}(1), P_\mathrm{SC}(2), \ldots, P_\mathrm{SC}(a)]$ obtained from SVM_M and SVM_C trained on motion and scene context descriptors, respectively, for action classes, i.e.,

$$
\begin{aligned}
&[P(1), P(2), \ldots, P(a)] \\
&\quad = [P_\mathrm{M}(1) P_C(1), P_\mathrm{M}(2) P_C(2), \ldots, P_C(a) P_\mathrm{M}(a)] .
\end{aligned}
$$

We use the fused probabilities as confidence to do the action classification.

7 Experiments and results

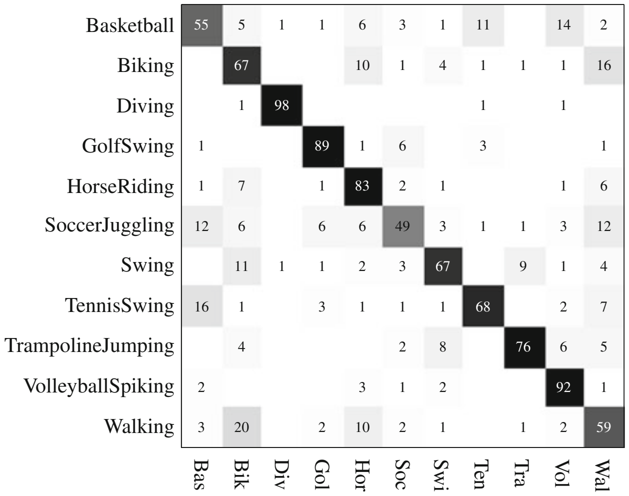
Experiments were performed on the following datasets: KTH, UCF11, UCF50, and HMDB51. The KTH dataset consists of six actions performed by 25 actors in a constrained environment, with a total of 598 videos. The HMDB51 dataset has 51 action categories, with a total of 6,849 clips. This dataset is further grouped into five types. In this dataset, general facial action type is not considered as articulated motion, which leaves the dataset with 47 action categories. The UCF11 dataset includes 1,100 videos and has 11 actions collected from YouTube with challenging conditions, such as low quality, poor illumination conditions, camera motions, etc. The UCF50 dataset has 50 actions with a minimum of 100 videos for each category, also taken from YouTube. This dataset has a wide variety of actions taken from different contexts and includes the same challenges as the UCF YouTube Action dataset.

In all of our experiments, we used three key frames from a single video to extract scene context features as explained before; however, we use all the frames in the video to get motion features without any pruning. We do not consider the audio, text, etc. contained in the video file to compute any of our features. Our method uses only the visual features. All the experiments have been performed under leave-one-out cross validation unless specified.

7.1 UCF11 dataset

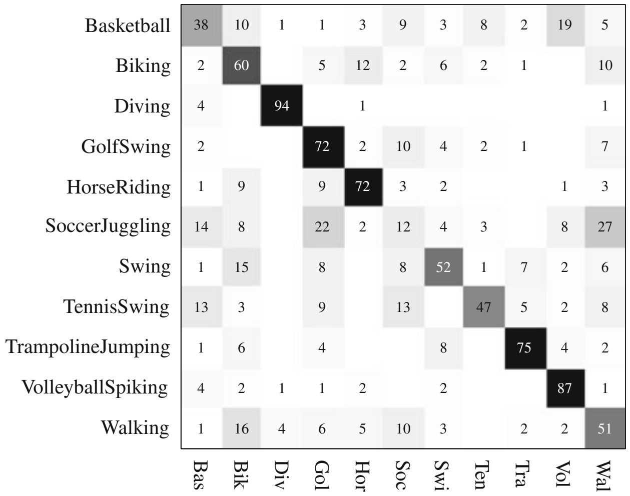
UCF11 is a very challenging dataset. We extract 400 cuboids of size 11 × 11 × 17 for the motion descriptor and a scene context descriptor from three key frames. We evaluate using leave-one-out cross validation. Our approach gives a performance of 73.20 % (Fig. 9), with a codebook of size 1,000. Motion descriptor alone gives a performance of 59.89 % (Fig. 10), and the scene context descriptor alone gives a performance of 60.06 % (Fig. 11). The idea of scene context plays a very important role in performing our approach. For example, the performance of motion descriptor for biking action is 49 %, and it has 21 % confusion with horse riding. After fusion with the scene context descriptor, which has 12 % confusion with horse riding, the performance increased to 67 % and the confusion with horse riding reduced to 10 %. The confusion decreased by 11 % and the performance increased by 18 %. This happens due to the complementary nature of probabilistic fusion where the individual strengths of the descriptors is preserved. This is also observed in "basketball" and "tennis swing" as shown in Fig. 9.

The performance reported by Liu et al. [11] using hybrid features obtained by pruning the motion and static features is 71.2 %. We performed ∼2 % better than Liu et al. [11]. Recently, Ikizler-Cinbis et al. [7] showed that their approach had 75.21 % performance, which was ∼2.1 % better than our approach. However, they performed computationally intense steps like video stabilization, person detector, and tracking, which were not done in our approach. By replacing the motion feature with MBH (4096-dimention codebook) [18] and following the exactly same experimental setup (SVM with a x^2 kernel) [18], the motion (MBH) and scene context descriptors gave us 83.13 and 46.57 %, respectively. When combined in multi-channel approach [18] gives 85.34 %, which is ∼1 % better than the best-known results on UCF11 as reported by Wang et al. [18].



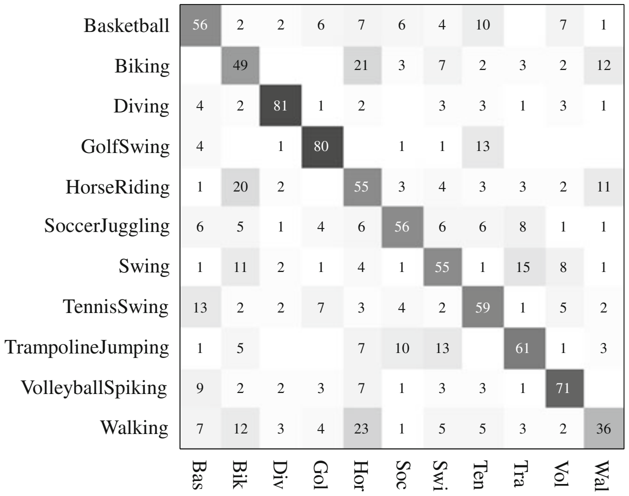
	Bas	Bik	Div	Gol	Hor	Soc	Swi	Ten	Tra	Vol	Wal
Basketball	55	5	1	1	6	3	1	11		14	2
Biking		67			10	1	4	1	1	1	16
Diving		1	98					1		1	
GolfSwing	1			89	1	6		3			1
HorseRiding	1	7		1	83	2	1			1	6
SoccerJuggling	12	6		6	6	49	3	1	1	3	12
Swing		11	1	1	2	3	67		9	1	4
TennisSwing	16	1		3	1	1	1	68		2	7
TrampolineJumping		4				2	8		76	6	5
VolleyballSpiking	2				3	1	2			92	1
Walking	3	20		2	10	2	1		1	2	59

Fig. 9 Confusion table for UCF11 dataset using the proposed framework i.e., probabilistic fusion of motion descriptor (dollar-gradient) and scene context descriptor. Average performance 73.20%



	Bas	Bik	Div	Gol	Hor	Soc	Swi	Ten	Tra	Vol	Wal
Basketball	38	10	1	1	3	9	3	8	2	19	5
Biking	2	60		5	12	2	6	2	1		10
Diving	4		94		1						1
GolfSwing	2			72	2	10	4	2	1		7
HorseRiding	1	9		9	72	3	2			1	3
SoccerJuggling	14	8		22	2	12	4	3		8	27
Swing	1	15		8		8	52	1	7	2	6
TennisSwing	13	3		9		13		47	5	2	8
TrampolineJumping	1	6		4			8		75	4	2
VolleyballSpiking	4	2	1	1	2		2			87	1
Walking	1	16	4	6	5	10	3		2	2	51

Fig. 11 Confusion table for UCF11 dataset using scene context descriptor. Average performance 60.06 %



	Bas	Bik	Div	Gol	Hor	Soc	Swi	Ten	Tra	Vol	Wal
Basketball	56	2	2	6	7	6	4	10		7	1
Biking		49			21	3	7	2	3	2	12
Diving	4	2	81	1	2		3	3	1	3	1
GolfSwing	4		1	80		1	1	13			
HorseRiding	1	20	2		55	3	4	3	3	2	11
SoccerJuggling	6	5	1	4	6	56	6	6	8	1	1
Swing	1	11	2	1	4	1	55	1	15	8	1
TennisSwing	13	2	2	7	3	4	2	59	1	5	2
TrampolineJumping	1	5			7	10	13		61	1	3
VolleyballSpiking	9	2	2	3	7	1	3	3	1	71	
Walking	7	12	3	4	23	1	5	5	3	2	36

Fig. 10 Confusion table for UCF11 dataset using motion descriptor (dollar-gradient). Average performance 59.89 %

7.2 UCF50 dataset

This is a very large and challenging dataset with 50 action categories. In this experiment, 1,000-dimension codebooks are used for both the motion and scene context descriptor. The individual performance of motion descriptor is 53.06 %; using our new scene context descriptor the performance is 47.56 %. After the fusion of both the descriptors, we have a performance of 68.20 %, which is a ∼15 % increase (Fig. 12).

The performance on rock climbing indoor using motion descriptors is 28; 11 % of the time it gets confused with rope climbing, and 10 % of the time rope climbing gets confused with rock climbing indoor. This is understandable because of the similar motion pattern in these actions. The performance of scene context descriptor for indoor rock climbing is 71 % with a confusion of 1 % with rope climbing, and the performance of rope climbing is 10 % with a confusion of 4 % with indoor rock climbing. Low confusion occurred because both the actions happened in two very different locations. Using our approach, we get 80 % performance on indoor rock climbing and 42 % performance on rope climbing. The complete confusion table is shown in Fig. 12. In some cases, the scene context descriptor performs badly compared to motion descriptor; for example, in bench press the performance using scene context is 54 % with 15 % confusion with pizza tossing. The reason for this is that both the actions are performed indoor in most cases. However, they have no confusion in motion descriptor. This increases the final performance of bench press to 71 %.

Figure 13 shows the performance by incrementally adding one action at a time from UCF50 to UCF11. The overall performance for the initial 11 actions using our approach is 70.56 %, and on all the 50 actions it is 66.74 %, a drop of 3.8 % in the overall performance in spite of adding 39 new actions. The fusion of both the descriptors consistently added 15.5 % to the motion descriptor with a variance of 1 and 17.3 % to the scene context descriptor with a variance of 9.3 % (Table 3).

It is interesting to note that substituting MBH (2048-dimension codebook) as the motion descriptor in the above experimental setup gave us the best performance of 76.90 %, where MBH and scene context descriptors gave 71.86 and 47.28 %, respectively.

7.3 HMDB51 dataset

The proposed approach has been tested on all the 51 categories in the HMDB51 dataset on original videos, and the

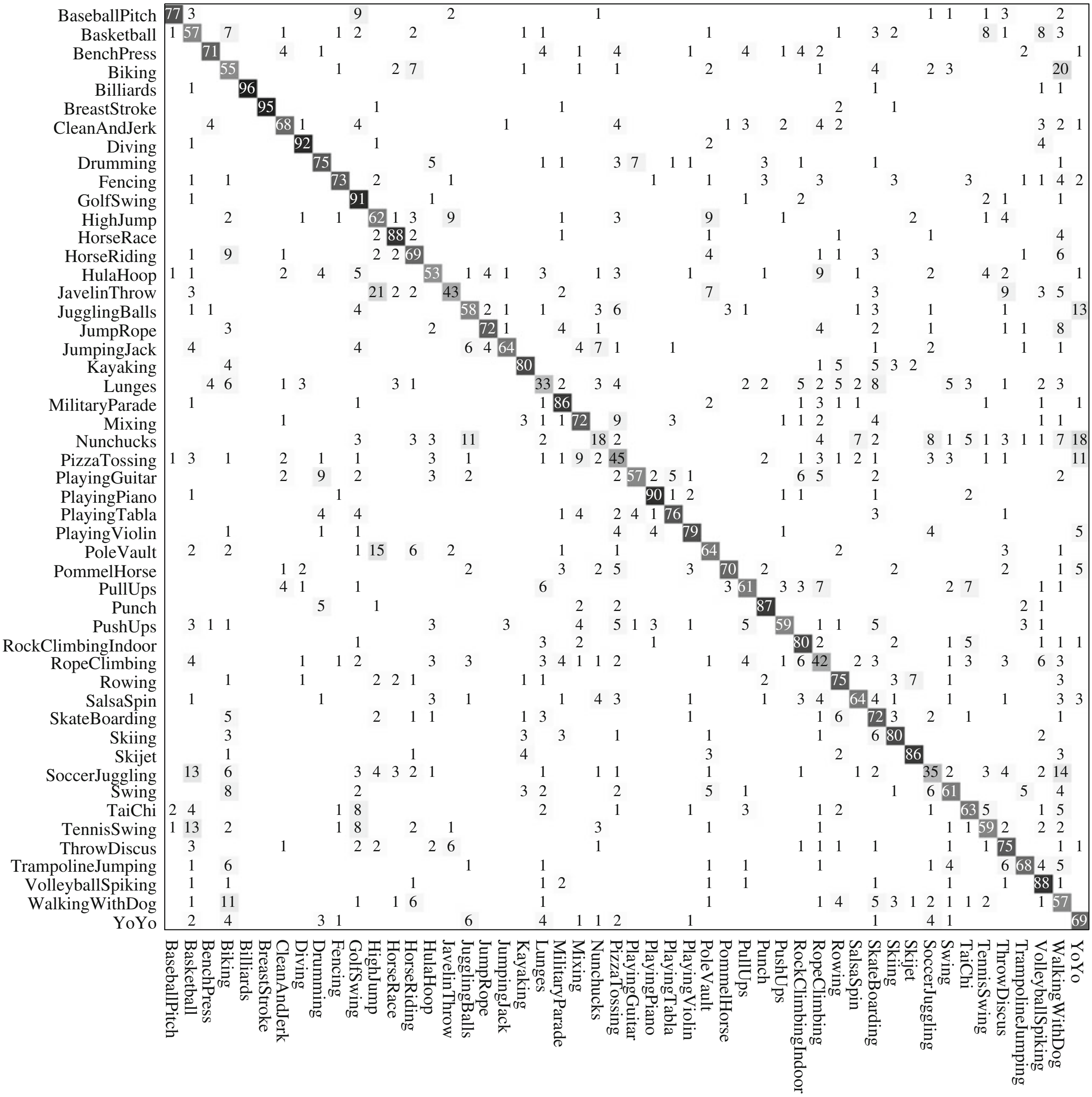


Fig. 12 Confusion table for UCF50 using our approach. Average performance 68.20 %

experimental setup was kept similar to [8] for comparison. We used the HOG/HOF features provided by the authors [8], which gave us 19.96 % for a codebook of size 2,000. The scene context descriptor is computed by extracting dense CSIFT on three key frames and quantizing using a codebook of size 2,000, which gave us 17.91 %. The proposed probabilistic fusion has 27.02 %, which is ~3.84 % higher than the best results reported by Kuehne et al. [8].

7.4 KTH dataset

We applied our proposed method on the KTH dataset. Although the idea of scene context is not useful in this dataset, experiments have been conducted simply to compare the performance of our method with other state-of-the-art results on the KTH dataset. The experimental setup is leave-one-out cross validation and a 1,000-dimension codebook is used. We

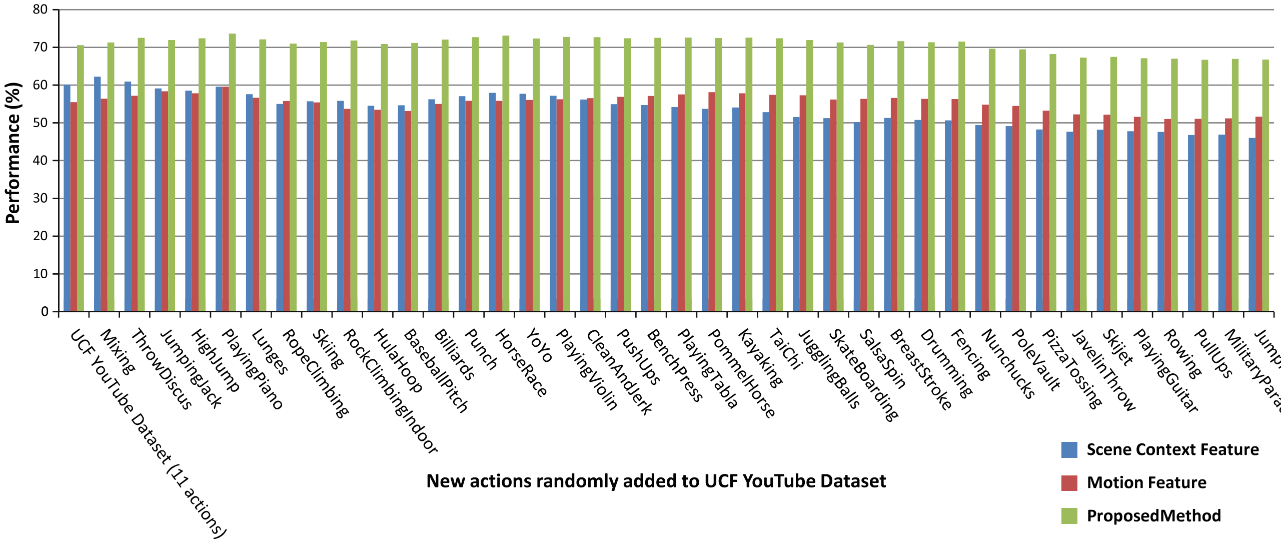


Fig. 13 Performance as new actions are added to UCF YouTube (UCF11) dataset from the UCF50 dataset

Table 3 Performance comparison on KTH dataset

Method	Acc (%)	Method	Acc (%)
Our method	89.79	Liu et al. [11]	91.3
Dollar et al. [4]	80.6	Wong et al. [22]	83.9

got a performance of 89.79 % using our approach, whereas scene context feature performance alone was 64.20 % and motion feature performance alone was 91.30 %. We had a 1.51 % drop in the performance due to the scene context features, in spite of the 25.95 % difference between scene context and motion features. This shows the robustness in performing the probabilistic fusion of both scene context and motion descriptors.

8 Conclusion

In this paper, we proposed an approach to perform recognition in large datasets like UCF50 and HMDB51. The proposed approach has the best performance on datasets like UCF11 (87.19 %), UCF50 (76.90 %), and HMDB51 (27.02 %). We showed that, as the number of categories increase, the motion descriptors become less discriminative. We also showed that the proposed scene context descriptor is more discriminative, and when properly fused with motion descriptors gives ∼15 and ∼4 % improvement on UCF50. Our approach does not require pruning of motion or static features, stabilization of videos, or detection and tracking of persons. The proposed method has the ability to do action recognition on highly unconstrained videos and also on large datasets.

References

1. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, 257–267 (2001)
2. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3273–3280 (2011)
3. Deng, J., Berg, A.C., Li, K., Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? In: Proceedings of the 11th European Conference on Computer Vision: Part V, pp. 71–84 (2010)
4. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 (2005)
5. Han, D., Bo, L., Sminchisescu, C.: Selection and context for action recognition. In: IEEE 12th International Conference on Computer Vision, pp. 1933–1940 (2009)
6. Hong, P., Huang, T.S., Turk, M.: Gesture modeling and recognition using finite state machines. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 410–415 (2000)
7. 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: Part I, pp. 494–507 (2010)
8. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: Proceedings of the International Conference on Computer Vision, pp. 2556–2563 (2011)
9. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
10. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1996–2003 (2009)
11. Liu, J., Shah, M.: Learning human actions via information maximization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)

12. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence, vol. 2, pp. 674–679 (1981)
13. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2929–2936 (2009)
14. van de Sande, K., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, pp. 1582–1596 (2010)
15. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th International Conference on Multimedia, pp. 357–360 (2007)
16. Snoek, C.G.M., Worring, M., Smeulders, A.W.M.: Early versus late fusion in semantic video analysis. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 399–402 (2005)
17. Song, Y., Zhao, M., Yagnik, J., Wu, X.: Taxonomic classification for web-based videos. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 871–878 (2010)
18. Wang., H., Klaser., A., Liu., C.L.: Action recognition by dense trajectories. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3169–3176 (2011)
19. Wang, Z., Zhao, M., Song, Y., Kumar, S., Li, B.: Youtubecat: learning to categorize wild web videos. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 879–886 (2010)
20. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. In: Computer Vision and Image Understanding, vol. 115, pp. 224–241 (2011)
21. Wilson, A., Bobick, A.: Parametric hidden markov models for gesture recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, pp. 884–900 (1999)
22. Wong, S.F., Kim, T.K., Cipolla, R.: Learning motion categories using both semantic and structural information. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–6 (2007)
23. Zheng, Y.T., Neo, S.Y., Chua, T.S., Tian, Q.: Probabilistic optimized ranking for multimedia semantic concept detection via rvm. In: Proceedings of International Conference on Content-Based Image and Video Retrieval, pp. 161–168 (2008)

Author Biographies

Kishore K. Reddy has a Bachelor of Technology degree in Electrical and Communication Engineering from Sree Vidhyanikethan Engineering College, Tirupathi, India and Master of Sciences in Electronic Systems and Engineering Management from Fachhochschule Südwestfalen, Soest, Germany. He received his Ph.D. degree in Electrical Engineering from the University of Central Florida in 2012. His research interests include gesture/action/activity recognition in videos, object detection, tracking and segmentation.

Dr. Mubarak Shah, Agere Chair Professor of Computer Science, is the founding director of the Computer Visions Lab at University of Central Florida (UCF). He is a co-author of three books (Motion-Based Recognition (1997), Video Registration (2003), and Automated Multi-Camera Surveillance: Algorithms and Practice (2008)), all by Springer. He has published extensively on topics related to visual surveillance, tracking, human activity and action recognition, object detection and categorization, shape from shading, geo registration, visual crowd analysis, etc. Dr. Shah is a fellow of IEEE, IAPR, AAAS and SPIE. In 2006, he was awarded the Pegasus Professor award, the highest award at UCF, given to a faculty member who has made a significant impact on the university. He is ACM Distinguished Speaker. He was an IEEE Distinguished Visitor speaker for 1997-2000, and received IEEE Outstanding Engineering Educator Award in 1997. He received the Harris Corporation's Engineering Achievement Award in 1999, the TOKTEN awards from UNDP in 1995, 1997, and 2000; SANA award in 2007, an honorable mention for the ICCV 2005 Where Am I? Challenge Problem, and was nominated for the best paper award in ACM Multimedia Conference in 2005 and 2010. At UCF he received Scholarship of Teaching and Learning (SoTL) award in 2011; College of Engineering and Computer Science Advisory Board award for faculty excellence in 2011;Teaching Incentive Program awards in 1995 and 2003, Research Incentive Award in 2003 and 2009, Millionaires' Club awards in 2005, 2006, 2009,2010 and 2011; University Distinguished Researcher award in 2007 and 2012. He is an editor of international book series on Video Computing; editor in chief of Machine Vision and Applications journal, and an associate editor of ACM Computing Surveys journal. He was an associate editor of the IEEE Transactions on PAMI, and a guest editor of the special issue of International Journal of Computer Vision on Video Computing. He was the program co-chair of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.