Efficient and effective human action recognition in video through motion boundary description with a compact set of trajectories

Jeong-Jik Seo, Jisoo Son, Hyung-Il Kim, Wesley De Neve, and Yong Man Ro

In: 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 1, 1-6, 2015.

To refer to or to cite this work, please use the citation to the published version:

Seo, J., Son, J., Kim, H., De Neve, W., and Ro, Y. M. (2015). Efficient and effective human action recognition in video through motion boundary description with a compact set of trajectories. *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition 1* 1-6. http://dx.doi.org/10.1109/FG.2015.7163123

Efficient and Effective Human Action Recognition in Video through Motion Boundary Description with a Compact Set of Trajectories

Jeong-Jik Seo[1], Jisoo Son[1], Hyung-Il Kim[1], Wesley De Neve[1,2], and Yong Man Ro[1,*]
[1]Image and Video Systems Lab, Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea
[2]Multimedia Lab, Ghent University-iMinds, Belgium

Abstract— **Human action recognition (HAR) is at the core of human-computer interaction and video scene understanding. However, achieving effective HAR in an unconstrained environment is still a challenging task. To that end, trajectory-based video representations are currently widely used. Despite the promising levels of effectiveness achieved by these approaches, problems regarding computational complexity and the presence of redundant trajectories still need to be addressed in a satisfactory way. In this paper, we propose a method for trajectory rejection, reducing the number of redundant trajectories without degrading the effectiveness of HAR. Furthermore, to realize efficient optical flow estimation prior to trajectory extraction, we integrate a method for dynamic frame skipping. Experiments with four publicly available human action datasets show that the proposed approach outperforms state-of-the-art HAR approaches in terms of effectiveness, while simultaneously mitigating the computational complexity.**

I. Introduction

Human action recognition (HAR) is one of the enabling technologies behind human-computer interaction and video scene understanding [1]. However, achieving effective HAR in an unconstrained environment is an open research challenge, given the frequent presence of background clutter, partial occlusions, viewpoint changes, and camera motion [1]. To overcome the aforementioned problems, and thus to realize effective HAR, discriminating a human action from background information is considered to be an important research task [1].

According to [1], local space-time feature extraction approaches using bag-of-words representations (e.g., 3-D Hessian [2], space-time interests points (STIP) [3], local trinary patterns (LTP) [4], Cuboids [5], and 3-D SIFT [6]) have shown promising levels of HAR effectiveness, mainly thanks to their robustness against partial occlusions and noise [1]. These approaches commonly focus on capturing edge and texture characteristics within 3-D space-time blocks defined by interest points. However, even when different types of motion are related to a human action within the 3-D space-time blocks used, the aforementioned approaches blend together the different types of motion, thus resulting in a loss of discriminative power [7].

To facilitate a more effective usage of motion information, trajectory-based feature extraction approaches have been proposed, following interest points along the temporal dimension with either a KLT-based tracker [7], SIFT matching [8], dense trajectory features (DTF) [9], or improved DTF (IDTF) [10]. By temporally tracking interest points, these approaches make it possible to automatically separate different types of motion information from background information. Therefore, as these approaches do not blend together different types of motion, they have been widely used for HAR. Among the different approaches mentioned, IDTF can be considered the state-of-the-art. IDTF extracts dense trajectories obtained by tracking uniformly sampled interest points using optical flow. In order to represent each trajectory, IDTF makes use of local descriptors such as histograms of oriented gradients (HOG) [11], histograms of optical flow (HOF) [12], motion boundary histograms (MBH) [13], and trajectory shapes (TS) [9]. Additionally, IDTF suppresses camera motion by estimating a homography, using a human detector to improve this estimation. That way, IDTF is able to more effectively represent the complicated motion of human actions. However, despite these strengths, IDTF still suffers from a number of weaknesses. First, the dense trajectories contain a substantial amount of redundancy [14]. Second, the extraction of dense trajectories is highly complicated due to the computation of optical flows and histogram-based descriptors (e.g., HOG, HOF, and MBH). Considering that real-world HAR requires both high levels of effectiveness and efficiency, it is necessary to address the two aforementioned weaknesses of IDTF.

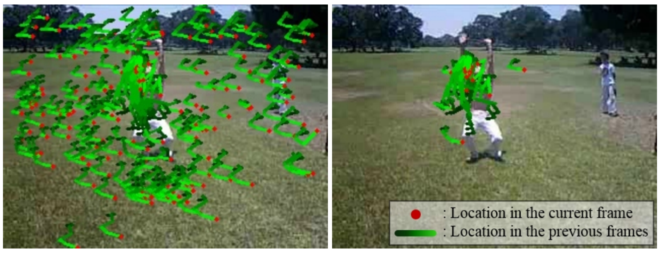
● : Location in the current frame
▬ : Location in the previous frames

Fig. 1. Visualization of dense trajectories after rejection by IDTF and after rejection by the proposed approach. Red dots indicate the interest point positions in the current frame, whereas the green curved lines denote the change in location of the interest point positions compared to the previous frame.

Motivated by the work of [14], we propose a method for motion boundary description with a compact set of trajectories, allowing for both effective and efficient HAR. In addition, to efficiently compute the optical flows needed for extracting trajectories, we propose to make use of a dynamic frame skipping technique, ignoring frames that contain less motion information. As such, we can summarize the main contributions of our paper as follows:

*Corresponding author (ymro@kaist.ac.kr)

978-1-4799-6026-2/15/$31.00 ©2015 IEEE

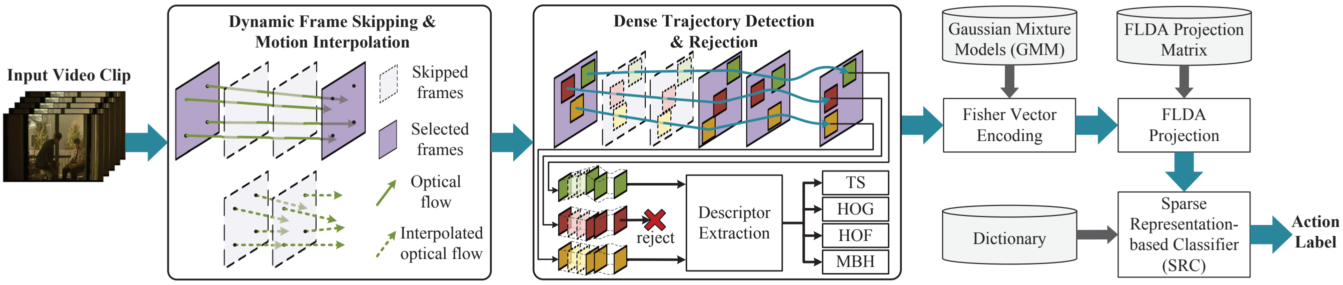
Input Video Clip
Dynamic Frame Skipping & Motion Interpolation
Skipped frames
Selected frames
Optical flow
Interpolated optical flow
Dense Trajectory Detection & Rejection
reject
Descriptor Extraction
TS
HOG
HOF
MBH
Gaussian Mixture Models (GMM)
Fisher Vector Encoding
FLDA Projection Matrix
FLDA Projection
Dictionary
Sparse Representation-based Classifier (SRC)
Action Label

Fig. 2. Overview of the proposed HAR framework.

- By removing redundant trajectories induced by camera motion, we are able to obtain a compact set of trajectories (please see Fig. 1). Since we are able to significantly mitigate the presence of trajectories irrelevant to human action, we can achieve higher levels of HAR effectiveness than previous research efforts.
- As we do not need to compute the optical flow for dynamically skipped frames, we are able to decrease the computational complexity of trajectory extraction.

Through comparative experimentation with four challenging human action datasets (i.e., HMDB51 [15], Hollywood2 [16], UCF50 [17], and UCF101 [18]), we show that the proposed approach allows improving HAR in terms of both effectiveness and efficiency.

The remainder of this paper is organized as follows. In Section II, we provide a high-level overview of the proposed approach. Next, in Section III, we explain our approach in more detail. In Section IV, we present our experimental results, demonstrating the efficiency and effectiveness of the proposed approach. Finally, we draw conclusions in Section V.

II. Overview of the Proposed HAR Framework

Figure 2 visualizes the proposed HAR framework. Our framework consists of three sequential modules: 1) dynamic frame skipping for efficient optical flow computation; 2) motion boundary information-based trajectory rejection and descriptor extraction along the trajectories obtained; and 3) human action classification.

For a given input video clip, we first select frames by making use of dynamic frame skipping (see the first step of Fig. 2). In order to allow for trajectory extraction, we then estimate the optical flow fields for the skipped frames. To that end, we simply interpolate between the optical flow fields computed for the frames retained, significantly reducing the overall complexity of optical flow computation. In the next step, for the optical flow fields obtained, we detect dense trajectories. To remove redundant trajectories caused by camera motion, we reject trajectories by making use of motion boundary information (see the second step of Fig. 2). Furthermore, the proposed method for trajectory rejection only retains the useful trajectories for describing a human action (that is, the proposed method only retains trajectories closely related to motion boundaries), facilitating both a reduction in computational complexity and a compact video content representation.

For classification purposes, we extract discriminative descriptors (HOG, HOF, MBH, and TS) from the selected trajectories. We further process the resulting descriptors by Fisher Vector (FV) encoding [19] using Gaussian Mixture Models (GMMs), an encoding approach that has recently shown state-of-the-art effectiveness. Finally, we generate an action label through the use of sparse representation-based classification (SRC) [20]. To satisfy the requirement of having an over-complete dictionary [20] and to increase the discriminative power, we adopt Fisher linear discriminant analysis (FLDA) [21] for the encoded FVs.

III. Proposed Trajectory-based Feature Extraction

In this section, we detail the proposed method for describing motion boundaries with a compact set of trajectories. In Section A, we first explain our motion boundary information-based method for trajectory rejection. Next, in Section B, we discuss our approach towards optical flow computation using dynamic frame skipping.

A. Motion Boundary Information-based Trajectory Rejection

According to [14], given the presence of redundant trajectories in a video clip (e.g., as caused by camera motion [22]), these non-relevant trajectories should be rejected in order to achieve more effective HAR. The authors of [10] identify four different types of trajectories in the IDTF framework that can be rejected: (1) static trajectories; (2) random trajectories; (3) trajectories with sudden and large displacements; and (4) trajectories affected by camera motion. According to [22], a trajectory caused by camera motion is highly redundant in nature, given that such a trajectory is irrelevant to the motion induced by a human action. In [10], the authors estimate a homography between adjacent frames to align their background. After background alignment using the homography and estimating the optical flow, if the maximum motion magnitude of a trajectory is lower than one pixel, they assume that the trajectory is generated by camera motion, and thus reject the trajectory. However, the homography cannot be exactly estimated in the case that there is a lack of pixel information in the background or in the case that pixel values are corrupted by noise or blur [10]. In this paper,

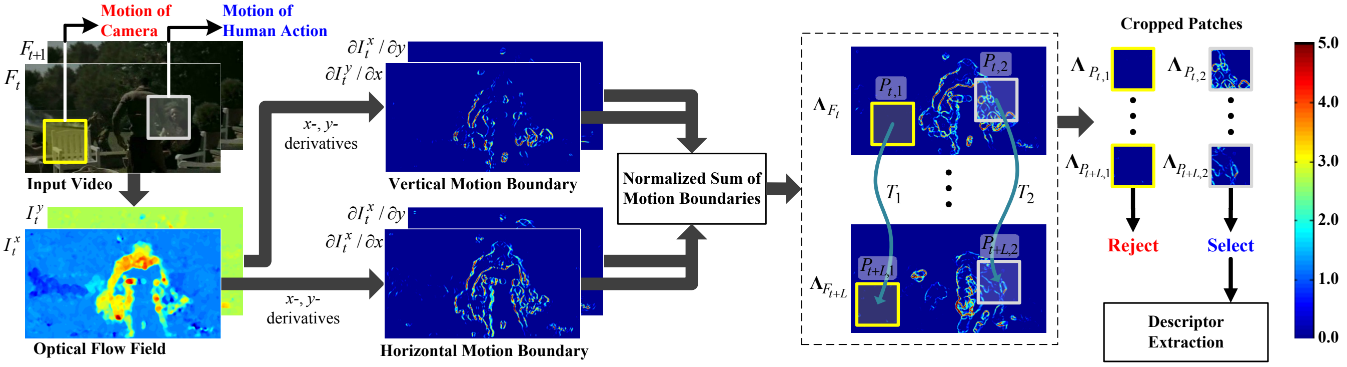

Fig. 3. Scheme for motion boundary information-based trajectory rejection. Using two consecutive frames (note that the yellow box delineates the camera motion, whereas the gray box delineates the human action), we estimate the x- and y- components of the corresponding optical flow field. Then, in order to measure the change of the motion vectors, we compute the partial derivatives of each component of the optical flow field (see the vertical and horizontal motion boundary). Next, we obtain the normalized sum of the motion boundaries. Depending on the motion boundary information of all the patches from T_1 and T_2, we decide whether or not a rejection is needed.

in order to reject redundant trajectories caused by camera motion, we propose a trajectory rejection approach based on motion boundary information. By leveraging motion boundary information, we can remove redundant trajectories more effectively than [10] (see Fig. 1 for a comparative example). In what follows, we provide more details regarding the aforementioned approach.

After having obtained the optical flow field $\omega_t \in \mathbb{R}^{(h \times w) \times 2}$ between two frames F_t at time t and F_{t+1} at time $t+1$ (h and w denote the height and width of the frames, respectively), we can separate it into two components (I_t^x, I_t^y), where $I_t^x \in \mathbb{R}^{h \times w}$ and $I_t^y \in \mathbb{R}^{h \times w}$ denotes the x- and y- component of the optical flow field, respectively. In order to compute the change of the motion vectors, we can calculate the partial derivatives $\partial I_t^x / \partial x$, $\partial I_t^x / \partial y$, $\partial I_t^y / \partial x$ and $\partial I_t^y / \partial y$. Each derivative captures the change of the optical flow (i.e., the motion boundary). Using the partial derivatives, we can compute the motion boundary information as follows:

$$\mathbf{\Lambda}_{F_t} = \frac{1}{4}\left(\left| \frac{\partial I_t^x}{\partial x} \right| + \left| \frac{\partial I_t^x}{\partial y} \right| + \left| \frac{\partial I_t^y}{\partial x} \right| + \left| \frac{\partial I_t^y}{\partial y} \right| \right), \quad (1)$$

where $\mathbf{\Lambda}_{F_t} \in \mathbb{R}^{h \times w}$ is the quantity that measures the motion boundary information of the frame F_t. Herein, Eq. (1) measures the normalized sum of the change of the motion vectors. Let $P_{t,j}$ denote a patch in the t-th frame F_t on the j-th trajectory T_j and let $\mathbf{\Lambda}_{P_{t,j}}$ denote the motion boundary information of the patch obtained from $\mathbf{\Lambda}_{F_t}$ by cropping the region of $P_{t,j}$. We can subsequently calculate the values λ_{F_t} and $\lambda_{P_{t,j}}$ by averaging all the elements of the matrices $\mathbf{\Lambda}_{F_t}$ and $\mathbf{\Lambda}_{P_{t,j}}$, respectively. Using λ_{F_t} and $\lambda_{P_{t,j}}$, we can then decide whether the patch $P_{t,j}$ contains substantial motion boundary information. For example, if $\lambda_{P_{t,j}}$ is larger than λ_{F_t}, then the patch $P_{t,j}$ is considered to contain more motion boundary information on average than any other region included in F_t. As shown on the right in Fig. 3, the patches from the human action region (as identified by the gray boxes) have plenty of information, while the patches from the camera motion region (as identified by the yellow boxes) are homogeneous. By counting the number of cases for which $\lambda_{P_{t,j}}$ is larger than λ_{F_t} for all patches from T_j, we can decide whether a trajectory should be rejected or not (e.g., if the total count is higher than a pre-defined threshold θ_{tr}, we let the trajectory T_j pass).

B. Dynamic Frame Skipping for Efficient Optical Flow Computation

A human action can be described by both motion and appearance information [1]. In order to extract motion information, optical flow has been widely used for the purpose of motion analysis [1]. According to [9], computing optical flow is the most time-consuming part of the dense trajectory extraction process for HAR. For the practical usage of HAR, reducing the complexity of optical flow computation is a crucial factor for obtaining a high efficiency.

Motivated by the observation that computing the optical flow between every two adjacent frames in a video clip is highly wasteful from a computational point-of-view, we propose a frame skipping-based method for improved efficiency. In particular, we adopted dynamic frame skipping, a technique that has been widely used for effective adjustment of the bit-rate in video transcoding [23]. According to [23, 24], dynamic frame skipping is based on the motion vectors obtained from the optical flow. In our framework, we modify the dynamic frame skipping method so that it is based on the difference between adjacent frames instead of making use of optical flow. That way, given that we only need to estimate the optical flow for selected frames and that the skipping scheme is only dependent on the difference operation, the computational complexity can be highly decreased.

Given two frames F_t and F_{t+i}, we first compute the absolute difference between these two frames in order to measure the significance of change:

$$C(x,y) = \begin{cases} 1, & |F_t(x,y) - F_{t+i}(x,y)| > \theta_p \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

where, $C(x,y)$ is a binary value that indicates whether the change between the two frames at pixel position (x,y) is higher than θ_p. Based on the value obtained from Eq. (2), we define the scene change index δ as a criterion for frame skipping:

$$\delta = \frac{1}{h \times w} \sum_{(x,y) \in F_t} C(x, y). \tag{3}$$

The normalized value defined in Eq. (3) reflects the significance of the total magnitude of motion change between two adjacent frames. If δ is larger than the pre-defined threshold value θ_n, we do not skip frame F_{t+i}, otherwise, we skip frame F_{t+i}. We limit the maximum number of frames to be skipped to five in order to avoid motion blurring. As a result, we only compute optical flow for the frames that have not been skipped.

Through the aforementioned frame skipping scheme, we can efficiently compute optical flow fields, given that we skip frames without significant changes in motion. However, in order to be able to further consider appearance and context information for the skipped frames when creating the discriminative descriptors, we need to devise a simple interpolation technique for obtaining representative motion fields for the skipped frames.

Assuming that the motion between two selected frames is linear in nature and that the number of skipped frames is denoted by s, we can estimate the optical flow between the selected frames by making use of the following procedure. Given that a motion vector between the selected frames passes through the skipped frames, we estimate the motion vector from the skipped frame to the next frame from the point where the motion vector passes through the skipped frame. We can then find the nearest pixel position around the pixel in the skipped frame from the path of the motion vector. Herein, to adjust the magnitude of the vector, we interpolate the motion vector at the pixel found by scaling the motion vector between the selected frames through $1/(s + 1)$. After interpolation, since the aforementioned forward mapping scheme generates a few holes, we apply a median filter as a post-processing step, filling the holes generated [25].

By skipping frames, interpolation errors are present, due to the degradation of the temporal resolution. However, we only skip a frame when the difference with the previous frame is not significant. In other words, additional motion information from such a skipped frame would not be a significantly important factor. Hence, this error does not critically influence the discriminative power, as also shown by our experimental results.

IV. Experiments

A. Experimental Setup

To evaluate the proposed approach, we used four challenging human action datasets that are publicly available: HMDB51 [15], Hollywood2 [16], UCF50 [17], and UCF101 [18]. HMDB51 contains 51 action classes, represented by 6,766 videos collected from movies and databases such as the Prelinger archive, YouTube, and Google Video. For the HMDB51 dataset, we followed the guidelines used by [15] (i.e., we generated three distinct training and testing splits from the database). Hollywood2 contains 12 action classes, represented by 1,707 videos obtained from 69 different movies. Herein, 823 videos are used for training and 884 videos are used for testing. In addition, UCF50 contains 50 action classes, represented by 6,676 videos [17]. These videos have been collected from YouTube. In this context, we made use of leave-one-group-out cross validation following the guidelines of [17]. UCF101, which is an extension of UCF50, contains 101 action classes, represented by 13,320 videos collected from YouTube [18]. For UCF101, we followed the original guide-lines outlined in [18]. For all datasets used, we measured the HAR effectiveness by means of average accuracy, with the exception of Hollywood2. For Hollywood2, we made use of mean average precision (mAP) [16].

In order to verify the effectiveness of the proposed approach, we compared our approach to IDTF [10] and other state-of-the-art approaches [14, 17, 22, 26-31]. In particular, for IDTF, we extracted histogram-based descriptors from volumes with a size of $N \times N \times L$ in the video frames, where N and L denote the spatial size of a volume and the trajectory length, respectively, and where we set these parameters to a value of 32 and 15, respectively [10]. Note that IDTF uses human detection for discriminating a human from the background [10]. In contrast, the proposed approach does not make use of human detection.

For all experiments, we empirically set the threshold values θ_{tr}, θ_p, and θ_n, as described in Section III, to 7, 5, and 0.02, respectively. We made use of the Farnebäck algorithm [32] to estimate the optical flow, conducting the polynomial expansion with a 7×7 Gaussian kernel that comes with a standard deviation of 1.5, and setting the maximum number of pyramid layers to 8. For Fisher vector encoding of the features extracted from the trajectories, we made use of principal component analysis (PCA) [33] to reduce the dimension of each descriptor, and we subsequently estimated GMMs by making use of 256 Gaussians. Note that we normalized the FVs through power and l_2-normalization [19]. To combine different types of descriptors into a single feature vector, we concatenated the normalized FVs. We subsequently applied FLDA [21] to increase the discriminative power as well as to reduce the dimensionality. Finally, we generated a human action label by making use of SRC [20].

Our evaluation consists of three experiments: 1) an experiment that investigates the effectiveness of the proposed approach; 2) an experiment that investigates the efficiency of the proposed approach; and 3) an experiment that compares the proposed approach with state-of-the-art HAR approaches [10, 14, 17, 22, 26-31]. We present our results in the following subsections.

B. Experiment 1: Evaluation of Effectiveness

To evaluate the effectiveness of the proposed approach, we performed experiments with the challenging HMDB51 and Hollywood2 datasets. For comparison purposes, we adopted IDTF [10], which is known as the most effective method

TABLE I

EFFECTIVENESS OF HAR (%) FOR IDTF AND THE PROPOSED APPROACH.

		TS	HOG	HOF	MBH	HOF+ MBH	ALL
HMDB51	IDTF	26.47	35.82	48.56	53.33	55.99	57.78
	Proposed	**32.61**	**36.19**	**49.15**	**54.88**	**56.51**	**58.91**
Hollywood2	IDTF	50.73	45.77	**60.82**	62.33	64.41	65.26
	Proposed	**51.98**	**48.72**	59.59	**63.84**	**65.39**	**65.37**

TABLE II

EFFICIENCY OF HAR (%) FOR IDTF AND THE PROPOSED APPROACH.

	Selected frames	Number of trajectories (trajectories/frames)	Speed(fps)
IDTF	29,258	128.72	7.82±0.003
Proposed	**18,319**	**90.41**	**9.84±0.004**

TABLE III

COMPARISON OF THE PROPOSED APPROACH TO THE STATE-OF-THE-ART.

HMDB51		**Hollywood2**	
Jain *et al.* 2013 [22]	52.10	Mathe *et al.* 2012 [14]	61.00
Wu *et al.* 2014 [26]	56.36	Jain *et al.* 2013 [22]	62.50
Cai *et al.* 2014 [27]	55.90	Jones *et al.* 2014 [29]	59.90
Narayan *et al.* 2014 [28]	58.70	Wang *et al.* 2013 [10]	64.30
Wang *et al.* 2013 [10]	57.20		
Proposed method	**58.91**	**Proposed method**	**65.37**
UCF50		**UCF101**	
Reddy *et al.* 2012 [17]	76.90	Karpathy *et al.* 2014 [31]	65.40
Ciptadi *et al.* 2014 [30]	90.50	Wu *et al.* 2014 [26]	84.16
Narayan *et al.* 2014 [28]	92.50	Cai *et al.* 2014 [27]	83.50
Wang *et al.* 2013 [10]	91.20	Wang *et al.* 2013 [34]	**85.90**
Proposed method	**93.70**	**Proposed method**	85.74

for feature extraction at the time of writing. In Table I, we present the difference in HAR effectiveness between IDTF and the proposed approach for a variety of descriptors. To facilitate a fair comparison, each framework made use of the same descriptors. Herein, the label ALL in Table 1 means the combination of all descriptors.

Overall, the proposed approach outperforms IDTF on the two datasets used, with the proposed approach achieving an average accuracy of 58.91% on HMDB51 and a mAP of 65.37% on Hollywood2 when making use of the combined descriptors. This can be mainly attributed to the proposed approach rejecting redundant trajectories, increasing the HAR effectiveness. Specifically, for the TS descriptor, the higher HAR effectiveness of the proposed approach stems from the fact that our approach is able to remove redundant trajectories caused by camera motion.

For the MBH descriptor, we could also obtain a gain in HAR effectiveness of about 1.6% on HMDB51 and 1.5% on Hollywood2. An MBH descriptor represents the orientation and the magnitude of the gradients of the optical flow. Therefore, an MBH descriptor extracted from a trajectory passing through a homogeneous motion region contains small values. Nevertheless, this descriptor is not negligible because all descriptors are normalized, thus dealt with equivalently. In addition, this descriptor is not discriminative because the gradients of the optical flow in the homogeneous motion region are nearly zero. Therefore, it adversely influences the effectiveness of HAR. On the other hand, since the trajectories from the homogeneous motion region were effectively removed, we could achieve a higher level of effectiveness with the MBH descriptor.

For Hollywood2, we can observe that the HAR effectiveness of the proposed approach is slightly lower than the HAR effectiveness of the IDTF framework using the HOF descriptor. This is mainly because of the properties of Hollywood2 and HOF. In the dataset, there are videos with a relatively large frame size and dominant human regions. In addition, HOF encodes the histogram of optical flow orientation. Therefore, a homogeneous motion region related to a human action can be suitably represented by HOF since that region contains human motion. However, our method typically removes a lot of trajectories that are detected in homogeneous motion regions related to human action. As a result, a limited amount of important information was lost, thus slightly decreasing the HAR effectiveness.

C. Experiment 2: Evaluation of Efficiency

To evaluate the efficiency of the proposed approach, we analyzed the skipped frame rate, the trajectory rejection rate, and the speed of operation (i.e., the number of frames per second (fps)). Herein, we measured the operation time from loading a video to obtaining all descriptors. In order to allow for a direct comparison, we used the 107 videos from the brush hair action class of HMDB51, having an average frame size of 335 × 240 pixels. All results were obtained on a desktop PC with an Intel Core i7-3770 (@3.40GHz) processor and 32 GB RAM, not using any parallel processing. We report our results by averaging the outcome of 10 test runs, in order to account for possible I/O and caching effects.

According to Table II, the proposed approach only selected 18,319 frames out of a total of 29,258 frames (about 63%), using the dynamic frame skipping method as described in Section II. The proposed approach selects 90.41 trajectories per frame. On the other hand, IDTF selects 128.72 trajectories per frame, which is larger than the number of trajectories per frame selected by the proposed approach. Additionally, we can observe that the processing speed of IDTF is about 7.82±0.003 fps. In contrast, the processing speed of the proposed approach is about 9.84±0.004 fps which is nearly a 20% increase in speed. Specifically, the proposed approach for optical flow computation led to a significant reduction in computing time of about 19%, compared to the computing time needed by DTF for optical flow computation. Furthermore, since the proposed approach does not make use of homography estimation and only extracts descriptors from the selected frames, the proposed approach is able to attain a lower computational complexity, in spite of the use of interpolated motion vectors for the skipped frames.

In summary, given the results presented in Table I and Table II, our approach was able to outperform the state-of-the-art IDTF approach in terms of both effectiveness and efficiency.

D. Comparison with the State-of-the-Art

To further verify the feasibility of the proposed approach, we compared our approach with a number of state-of-the-art HAR approaches [10, 14, 17, 22, 26-31], using the four datasets previously described. To facilitate a fair comparison, we used the descriptor that combines the four types of descriptors discussed earlier. As can be seen in Table III, for all datasets, the proposed approach achieves better or comparable levels of HAR effectiveness. In the context of UCF101, the method of Wang [34] achieved the highest effectiveness. However, the effectiveness was accomplished by human detection and spatio-temporal pyramids [34], winning the THUMOS challenge [35]. When spatio-temporal pyramids are not applied, then the average accuracy lowers to 84.8%, which is lower than the effectiveness of the proposed approach.

V. Conclusions

In this paper, we proposed a trajectory rejection method, with the aim of improving the effectiveness of HAR. At the same time, in order to mitigate the computational complexity of optical flow computation prior to extracting trajectories, we integrate a method for dynamic frame skipping. Through experimentation with four publicly available datasets, we demonstrated that the proposed approach outperforms state-of-the-art approaches in terms of both effectiveness and efficiency. In particular, at a gain of about 20% in computational efficiency, our approach was still able to achieve comparable and reliable levels of HAR effectiveness.

VI. Acknowledgements

This work was supported by the ICT R&D program of MSIP/IITP. [13912-06006, Development of the Filtering Technology for Objectionable Streaming Contents on Smart Platform].

References

[1] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976-990, 2010.

[2] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *European Conf. Computer Vision*, 2008, pp. 650-663.

[3] I. Laptev, "On space-time interest points," *Int'l Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107-123, 2005.

[4] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in *IEEE Int'l Conf. Computer Vision*, 2009, pp. 492-497.

[5] P. Dollr, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *IEEE Int'l Works.* Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65-72.

[6] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in *Int'l Conf. Multimedia*, 2007, pp. 357-360.

[7] P. Matikainen, M. Hebert, and R. Sukthankar, "Trajectons: Action recognition through the motion analysis of tracked features," in *IEEE Int'l Conf. Computer Vision Works.*, 2009, pp. 514-521.

[8] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 2004-2011.

[9] H. Wang, A. Klser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int'l Journal of Computer Vision*, vol. 103, no. 1, pp. 60-79, 2013.

[10] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE Int'l Conf. Computer Vision*, 2013, pp. 3551-3558.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 886-893.

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1-8.

[13] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conf. Computer Vision*, 2006, pp. 428-441.

[14] S. Mathe and C. Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in *European Conf. Computer Vision*, 2012, pp. 842-856.

[15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *IEEE Int'l Conf. Computer Vision*, 2011, pp. 2556-2563.

[16] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 2929-2936.

[17] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971-981, 2012.

[18] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.

[19] F. Perronnin, J. Snchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *European Conf. Computer Vision*, 2010, pp. 143-156.

[20] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210-227, 2009.

[21] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[22] M. Jain, H. Jgou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 2555-2562.

[23] J.-N. Hwang, T.-D. Wu, and C.-W. Lin, "Dynamic frame-skipping in video transcoding," in *IEEE Works. Multimedia Signal Processing*, 1998, pp. 616-621.

[24] C.-Y. Chen, C.-T. Hsu, C.-H. Yeh, and M.-J. Chen, "Arbitrary frame skipping transcoding through spatial-temporal complexity analysis," in IEEE Region 10 Conf. TENCON, 2007, pp. 1-4.

[25] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, et al., "Real-time visibility-based fusion of depth maps," in *IEEE Int'l Conf. Computer Vision*, 2007, pp. 1-8.

[26] J. Wu, Y. Zhang, and W. Lin, "Towards good practices for action video encoding," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 2571-2578.

[27] Z. Cai, L. Wang, X. Peng, and Y. Qiao, "Multi-View Super Vector for Action Recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 596-603.

[28] S. Narayan and K. Ramakrishnan, "A Cause and Effect Analysis of Motion Trajectories for Modeling Actions," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 2633-2640.

[29] S. Jones and L. Shao, "A Multigraph Representation for Improved Unsupervised/Semi-supervised Learning of Human Actions," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 820-826.

[30] A. Ciptadi, M. S. Goodwin, and J. M. Rehg, "Movement Pattern Histogram for Action Recognition and Retrieval," in *European Conf. Computer Vision*, 2014, pp. 695-710.

[31] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1725-1732.

[32] G. Farnebck, "Two-frame motion estimation based on polynomial expansion," in *13th Scandinavian Conf. Image Analysis*, 2003, pp. 363-370.

[33] I. Jolliffe, 2005. *Principal Component Analysis*: Wiley Online Library.

[34] H. Wang and C. Schmid, "LEAR-INRIA submission for the THUMOS workshop," in *ICCV Works. Action Recognition with a Large Number of Classes*, 2013.

[35] [Online] Available: http://crcv.ucf.edu/ICCV13-Action-Workshop/