3D GLOH Features for Human Action Recognition

Ashwan Abdulmunem
School of Computer Science and Informatics
Cardiff University, UK
School of Science
University of Kerbala, Iraq
Email: AbdulmunemAA@cardiff.ac.uk

Yu-Kun Lai
School of Computer Science and Informatics
Cardiff University, UK
Email: LaiY4@cardiff.ac.uk

Xianfang Sun
School of Computer Science and Informatics
Cardiff University, UK
Email: SunX2@cardiff.ac.uk

***Abstract*—Human action recognition from videos has wide applications and has attracted significant interests. In this work, to better identify spatio-temporal characteristics, we propose a novel 3D extension of Gradient Location and Orientation Histograms, which provides discriminative local features representing not only the gradient orientation, but also their relative locations. We further propose a human action recognition system based on the Bag of Visual Words model, by combining the new 3D GLOH local features with Histograms of Oriented Optical Flow (HOOF) global features. Along with the idea from our recent work to extract features only in salient regions, our overall system outperforms existing feature descriptors for human action recognition for challenging real-world video datasets.**

I. Introduction

Human action recognition is one of the most important topics in the video processing domain. Analysis of human actions in videos is a crucial problem in computer vision because many applications are dependent on it, e.g., human-computer interaction, content-based video retrieval, visual surveillance, analysis of sport events, video manipulation, etc. Recognising human actions from video is also a very challenging problem because of the fact that the physical body of a human subject doing the same action can look very different depending on the situation. For instance, similar actions of the same person with different clothes or in different illumination and background can result in large appearance variation. The same action performed by two different people may look quite dissimilar in many ways as well.

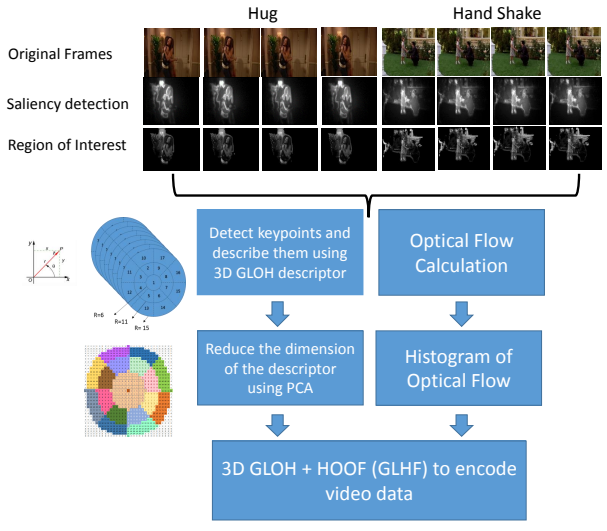

Figure 1. Feature extraction using the S-GLHF (Saliency Guided 3D GLOH and HOOF) descriptor in our action recognition system.

To cope with such variation, extracting features from video frames is typically employed as an essential component in an action recognition system. If the extracted features are informative and selective, then accurate and efficient action recognition can be achieved. Feature extraction can be achieved either locally or globally, and a combination of complementary features can often produce more robust and informative features for encoding video information. Existing feature extraction methods for human action recognition can mainly be categorised into motion-based, shape-based and texture-based methods, based on different image properties being considered. In motion-based methods, the optical flow features are used to represent the motion of action in the video frames [1]. In shape-based approaches, an action is encoded as a shape descriptor to preserve the action information [2], [3]. In texture-based methods, selective and robust local texture information within some interest regions in the video frames is extracted to encode the action [4], [5], [6]. As an alternative, recent work uses deep Convolutional Neural Networks (CNNs) [7], [8] to learn features and perform classification, which avoids handcrafted features. Such methods benefit from a large number of training data and produce competitive results, although the learnt features can be less intuitive to interpret. This paper focuses on designed features which are generally easier to implement, and do not require a large training set.

2D descriptors have achieved notable success in object detection and recognition. Considering videos as 3D spatio-temporal volumes, efforts have been made to extend 2D descriptors in the image domain to 3D volumes in the video domain. Recent work has demonstrated that such descriptors can represent the video information more effectively than those with 2D features. Klaser et al. [4] described the video information as a 3D histogram of gradients with different scales and locations in the spatio-temporal domain. Willems et al. [9] introduced ESURF by extending the SURF descriptor to 3D patches. Zhang et al. [6] introduced a 3D descriptor to capture

video information, which is called simplex-based orientation decomposition (SOD). This descriptor represents video data based on decomposing the orientations of visual cues into three angles, and then transforming the decomposed angles into the simplex space. Scovanner et al. [5] extended the SIFT descriptor to the spatio-temporal domain. Although features such as 3D SIFT are effective in representing the local spatio-temporal characteristics, they are essentially histograms of (gradient) distribution in the local neighbourhood of selected interest points. The spatial location of such distribution is ignored. Such information, however, can be discriminative if the video contains spatially varying details, which is common for real-world videos.

In this paper, we propose a novel effective feature called 3D GLOH (Gradient Location and Orientation Histogram), which describes local spatially varying information for video data. It detects interest points in the video and then describes them in 3D log-polar coordinates. This descriptor is an extension of the 2D GLOH descriptor [10] and we will demonstrate that it better captures the characteristics of local video information than existing features. Moreover, we propose an action recognition system, that uses 3D GLOH for extracting local features, along with histograms of oriented optical flow (HOOF) [11] for extracting global features. We further employ the idea from our recent work [12] and extract features only in salient regions for action recognition. We evaluate the new combined descriptor using a variety of video datasets. The new descriptor outperforms the state-of-the-art descriptors for challenging real-world videos with uncontrolled complicated environment, such as the UCF-Sports and TV-Human Interaction datasets.

The main contributions of the paper can be summarised as follows:

1) We propose a novel 3D GLOH feature and demonstrate its usefulness for human action recognition.
2) We develop a novel combination of local and global descriptors, which outperforms existing descriptors in action recognition with challenging real-world videos.

II. Proposed Method

The overall framework of our human action recognition system encodes video sequences using a combined local and global representation, along with the Bag of Visual Words (BoVW) framework. The local features are represented by our proposed 3D GLOH from only salient regions in the video frames [12] and the global features are represented using Histograms of Oriented Optical Flow (HOOF). Figure 1 illustrates the main steps of the proposed system for feature extraction. We now describe the system with an emphasis on the novel 3D GLOH descriptor as follows.

A. 3D Gradient Location and Orientation Histograms (3D GLOH)

To capture the gradient distribution and localise it in the neighbouring spatio-temporal domain, we extend the GLOH descriptor proposed by Milkolajczky and Schmid [10] to 3D in a log-polar location partitioning. More specifically, we first

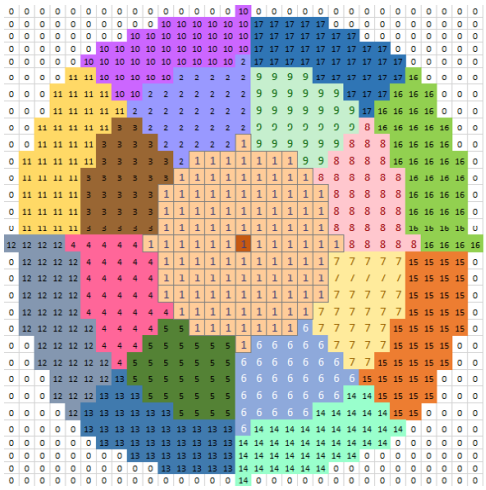

Figure 2. The neighbourhood local region labelling at an interest point used for computing the GLOH descriptor in a log-polar domain.

detect interest points using the standard 2D SIFT [13]. For each detected interest point, we consider its neighbourhood as a cylinder in the spatio-temporal volume, with a diameter of 31 pixels in the spatial domain and a height of 8 pixels (frames) along the temporal domain. The cylinder is further divided in both the spatio- and temporal domains to provide localised distribution. In the temporal domain, the cylinder is split into two halves each with 4 frames. In the spatial domain, following [10] a log-polar location grid is used with three bins in the radial direction (with the radii set to 6, 11 and 15) and 8 in the angular direction for each slice, which results in 17 location bins (see Figure 2), where the central bin is not divided in angular directions. Cartesian coordinate system is transformed into the polar coordinate system through the following equations:

$$r_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}, \tag{1}$$

$$\theta_i = \tan^{-1}((y_i - y_c)/(x_i - x_c)), \tag{2}$$

where (x_i, y_i) is the coordinate of pixel in the Cartesian coordinate system, (r_i, θ_i) is the radius and the angle in the polar coordinate system. (x_c, y_c) is the coordinate of the interest point. This leads to $17 \times 2 = 34$ local regions in the spatio-temporal domain.

For each pixel in a local region, 3D gradients are calculated, similar to 3D SIFT [5]. The 3D gradient orientation for each pixel are described using two angles θ and ϕ, which are defined as follows:

$$\theta(x, y, t) = \tan^{-1}(L_y/L_x), \tag{3}$$

$$\phi(x, y, t) = \tan^{-1}(L_t/\sqrt{L_x^2 + L_y^2}), \tag{4}$$

where L is the intensity of the video frame, L_x, L_y, and L_t are the intensity gradients w.r.t. x, y and time, computed respectively using finite difference approximations: $L(x + 1, y, t) - L(x - 1, y, t)$, $L(x, y + 1, t) - L(x, y - 1, t)$ and

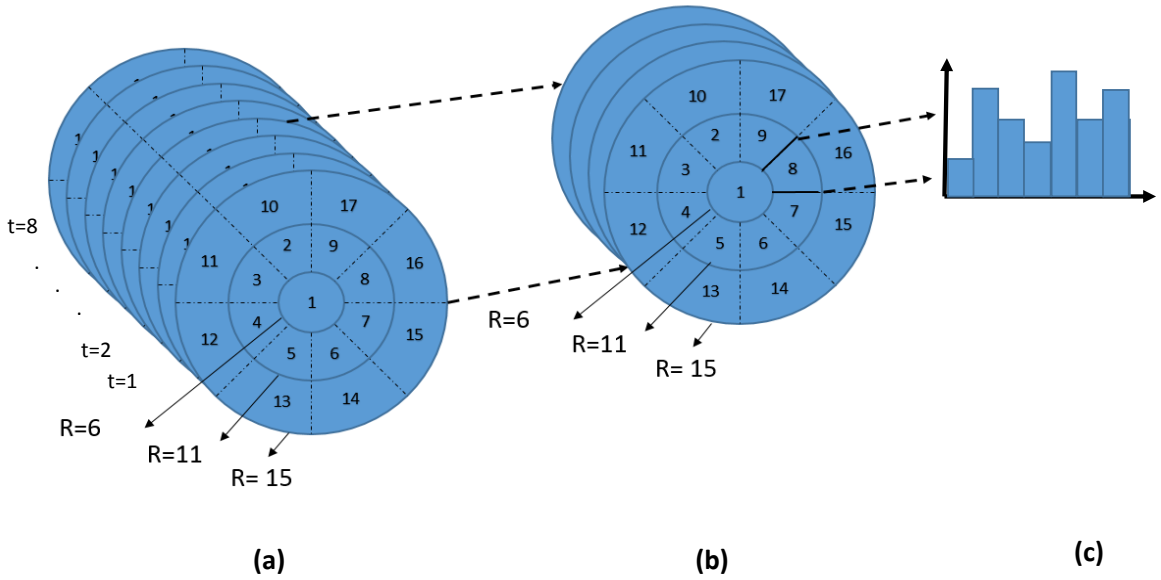

Figure 3. 3D GLOH representation: a) Neighbourhood of the interest point as a cylinder with a diameter of 31 and 8 frames in the spatio-temporal domain. b) Histogram computation over local regions with spatial domain split into 17 log-polar location grid and temporal domain split into two halves. c) Histogram of a local region.

$L(x, y, t+1) - L(x, y, t-1)$. θ and ϕ encode the angles for the 3D gradient direction.

Each gradient orientation angle is quantised into N bins (by default we use $N = 16$). As two angles are used to describe a 3D orientation, the descriptor is a vector of $2N \times 34$ dimension as shown in Figure 3.

The resulting descriptor is high dimensional, which makes computation expensive. For example, when the default $N = 16$ is used, the histogram dimension is $2 \times 16 \times 34 = 1088$. We use Principal Component Analysis (PCA) to reduce the dimensionality. The covariance matrix for PCA is estimated using the training examples in the datasets, and 192 dominant eigenvectors are used to reduce the dimension to the same level as SIFT features.

B. Human Action Recognition using S-GLHF Descriptor

As we will show later, our proposed 3D GLOH descriptor is particularly effective in describing local spatio-temporal distribution at each interest point. Following our recent work [12], we apply saliency detection [14] to identify salient (foreground) objects from each video frame, and only consider keypoints in the foreground. This helps suppress the impact of spurious keypoints by incorporating some "semantic" information. The 3D GLOH descriptor is then complemented with a global descriptor namely Histograms of Oriented Optical Flow (HOOF) [11], which produces a histogram representing each frame of the video.

To summarise the characteristics of the whole video, we employ a Bag-of-Visual-Words framework. We build a vocabulary of visual words for each of the two descriptors (3D GLOH and HOOF) using k-means clustering of features extracted from all the training videos in the dataset. 2000 visual words are used for each descriptor as it gives a right balance of efficiency and performance. Once this is done, each feature vector is mapped to the closest visual word in the vocabulary. For each video a feature vector is obtained by concatenating two histograms measuring the distribution of visual words in the video. This combined descriptor (which we call GLHF) benefits from local and global representations to describe the information of the video in an informative and selective manner. For classification, we use multi-class kernel SVM classifier using Radial Basis Function (RBF) kernels. The SVM kernel parameters are automatically optimised using grid search with 5-fold cross validation.

III. Experimental Results

In this section we report results on several benchmark datasets, and discuss how our method behaves with varying key parameters.

A. Datasets

To investigate the effectiveness of our approach for action recognition, experiments were conducted on three video datasets, namely UCF Sports [15], TV-Human Interaction [16] and KTH [17]. These datasets differ in several aspects such as recording conditions, scenarios, number of actors in video.

The UCF-Sports dataset contains 10 different sport action categories (Diving, Horse-Riding, Kicking, Swinging, Lifting, Walking, Running, Skating, Golf, High-bar), recorded in real-world environment.

Figure 4. Benchmark datasets used to evaluate our method. Top to bottom: images from videos in UCF-sports, TV-Human Interaction and the KTH datasets.

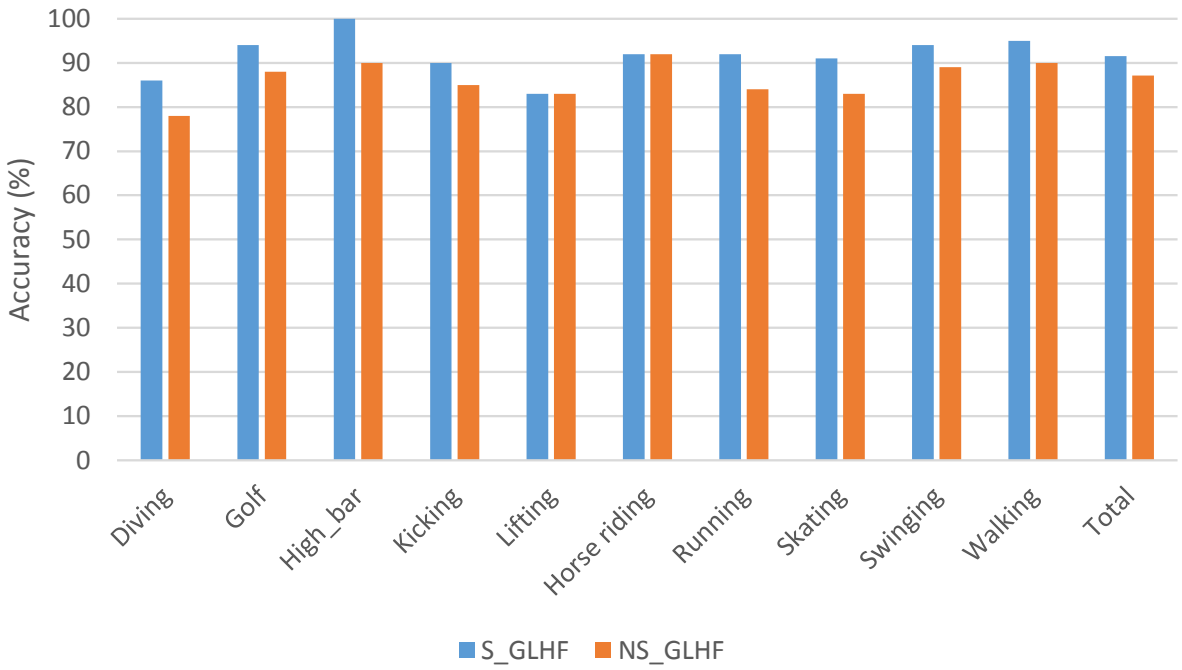

Figure 5. The recognition rates of The UCF Sport dataset for each individual action and the total accuracy with saliency guidance (S-GLHF) and without saliency guidance (NS-GLHF).

The TV-Human Interaction dataset has been collected from different movies and includes 5 action classes (Handshake, Highfive, Hug, Kiss, and Negative where Negative action does not contain any interaction). It contains 300 videos, 50 for each interaction action class and 100 for the negative class.

The KTH dataset includes 6 action classes (Boxing, Hand-clapping, Handwaving, Jogging, Running, Walking). In total, there are 600 videos in this dataset. Each action was performed by 25 actors, and each person has 4 records for each action in a *controlled* environment. Figure 4 shows some examples for each dataset.

The standard test setup was used (training/test separation for KTH and TV-Human Interaction, and leave-one-out testing for UCF-Sports) to allow fair comparison with prior work.

B. Results and discussions

We performed extensive experiments using these standard datasets to study the effectiveness of our proposed 3D GLOH

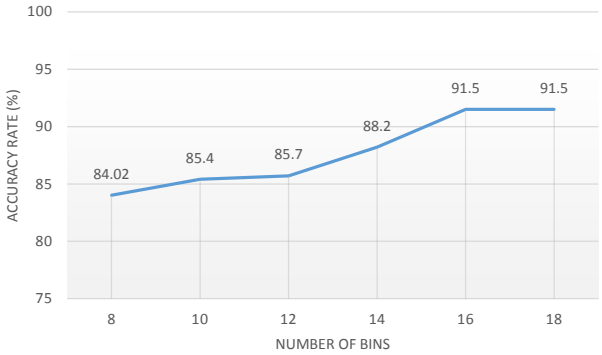

Figure 6. Recognition rate using different numbers of bins for the 3D-GLOH descriptor.

	Di	Go	HB	Ki	Lf	HR	Rn	Sk	Sw	Wa
Diving	0.85	0	0	0.15	0	0	0	0	0	0
Golf	0	0.94	0	0	0	0	0	0.06	0	0
HB	0	0	1.0	0	0	0	0	0	0	0
Kicking	0	0	0	0.90	0	0	0	0.05	0	0.05
Lifting	0	0	0	0	0.83	0	0	0	0.17	0
HR	0	0.09	0	0	0	0.91	0	0	0	0
Running	0	0	0	0	0	0	0.92	0	0	0.08
Skating	0.09	0	0	0	0	0	0	0.91	0	0
Swinging	0	0	0	0	0.06	0	0	0	0.94	0
Walking	0	0	0	0	0	0	0.05	0	0	0.95

Figure 7. Confusion matrix for The UCF-Sport dataset with our action recognition system. HB (High bar), HR (Horse Riding).

descriptor and the human action recognition system.

For the UCF Sports dataset, Figure 5 shows the performance of recognising each class of videos using 3D GLOH and HOOF descriptors. The method works consistently well in all categories, and in particular by using the saliency guidance, the recognition rate increases for every class of videos (blue bars with saliency vs. orange bars without saliency). A key parameter in the 3D GLOH descriptor is the number of bins N when histograms are built. To investigate the behaviour of our method with changing N, results are reported in Figure 6, and it can be seen that $N = 16$ achieves good results. Thus unless for comparative purpose, we use this setting for all the experiments in the paper. The confusion matrix of the results obtained using our system is reported in Figure 7. We compare our method with state-of-the-art methods which reported the performance on the UCF Sport dataset (see Table I). Our method (S-GLHF) outperforms state-of-the-art methods. The improvement 0.6% is still quite significant as the performance has already been over 90%.

The TV-Human Interaction dataset is more complicated as it involves interactions between multiple subjects. Figure 8 shows the recognition rate of our approach for each individual action. The performance is consistently good, especially with saliency guidance. Compared with existing methods tested on this dataset (see Table II), our method achieves 75.3% accuracy, which improves the state-of-the-art method [18] (66.1%) by a significant margin (9.2%). In fact, for each action, our method achieves over 70% accuracy, which is better than the average performance of [18].

Our 3D GLOH feature exploits the spatio-temporal distribution of gradients to provide a more discriminative descriptor. As a result, our 3D GLOH feature may not be very effective if the video data contains little texture. An example of such kind of data is the KTH dataset (cf. Figure 4). This dataset is relatively easy as it has a clean background and was captured in a controlled environment. However, the images are relatively low-resolution and do not contain much texture. Our method achieves 94.9% accuracy, which is close to some of the recent methods [6] (94.8%) but not as good as [12] which achieves 97.2%. Nevertheless, for more challenging real-world datasets, we have shown that the proposed 3D GLOH descriptor is effective and outperforms existing methods.

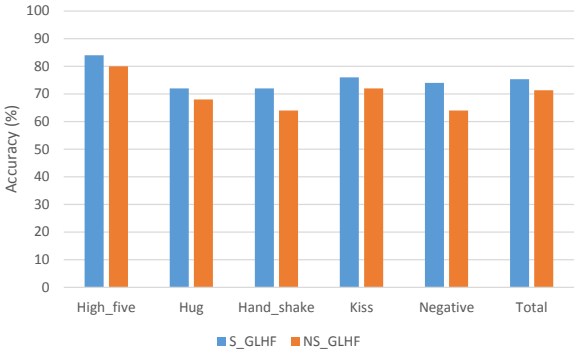

Figure 8. The recognition rate of the TV-Human Interaction dataset for each individual action with saliency (S_GLHF) and without saliency (NS_GLHF).

	HF	Hug	HS	KS	Neg
HF	0.84	0.08	0.08	0	0
Hug	0	0.72	0.12	0.08	0.08
HS	0.08	0.12	0.72	0.08	0
KS	0	0.12	0.08	0.76	0.04
Neg	0.04	0.08	0.04	0.1	0.74

Figure 9. Confusion matrix for TV-Human Interaction dataset using our method (with saliency guidance). HF (High Five), HS (Hand Shake), KS (Kiss) and Neg (Negative).

IV. Conclusion

In this paper, we introduce a new local descriptor for video data namely 3D GLOH and propose a human action recognition system using the proposed local descriptor along with a global descriptor. The 2D GLOH descriptor is extended to video frames by partitioning the cylindrical

Table I
RECOGNITION ACCURACY COMPARISONS ON THE UCF SPORTS DATASET.

Methods on (UCF-Sports)	Accuracy (%)
Raptis [19]	79.4
Ma [20]	81.7
Kalser [21]	85.0
Everts [22]	85.6
Le [23]	86.5
Zhang [6]	87.5
Wang [24]	88.0
Ma [25]	89.4
Abdulmunem [12]	90.9
Our Method (S-GLHF)	**91.5**

Table II
RECOGNITION ACCURACY COMPARISONS ON THE TV-HUMAN INTERACTION DATASET.

Methods on (TV-Human Interaction)	Accuracy (%)
Patron-Perez [16]	32.8
Yu [26]	56.0
Gaidon [27]	62.4
Yu [18]	66.1
Our Method (S-GLHF)	**75.3**

local neighbourhood of an interest point into spatio-temporal bins and calculating 3D histograms of gradients in local bins. The experimental results show that the proposed 3D GLOH descriptor is effective in capturing localised spatio-temporal information and the overall system outperforms the state-of-the-art methods in terms of recognition accuracy for challenging real-world datasets including UCF-Sport and TV-Human Interaction datasets.

Feature descriptors are widely used for video analysis. The proposed 3D GLOH descriptor can be useful for analysing videos, especially for those with rich textures. We would like to investigate its effectiveness in other video analysis applications.

ACKNOWLEDGMENT

This work was funded by the Iraqi Ministry of Higher Education and Scientific Research (MHESR).

REFERENCES

[1] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *ECCV*, vol. 3952, 2006, pp. 428–441.

[2] J. Niebles and F.-F. Li, "A hierarchical model of shape and appearance for human action classification," in *CVPR*, 2007, pp. 1–8.

[3] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[4] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *BMVC'08*, 2008, pp. 995–1004.

[5] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in *Proceedings of the 15th International Conference on Multimedia*. New York, NY, USA: ACM, 2007, pp. 357–360.

[6] H. Zhang, W. Zhou, C. Reardon, and L. Parker, "Simplex-based 3D spatio-temporal feature description for action recognition," in *CVPR*, 2014, pp. 2067–2074.

[7] Q. Li, Z. Qiu, T. Yao, T. Mei, Y. Rui, and J. Luo, "Action recognition by learning deep multi-granular spatio-temporal video representation," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '16, 2016, pp. 159–166. [Online]. Available: http://doi.acm.org/10.1145/2911996.2912001

[8] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *CVPR*, 2015, pp. 4305–4314.

[9] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *ECCV'08*, 2008, vol. 5303, pp. 650–663.

[10] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[11] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *CVPR*, 2009, pp. 1932–1939.

[12] A. Abdulmunem, Y.-K. Lai, and X. Sun, "Saliency guided local and global descriptors for effective action recognition," *Computational Visual Media*, vol. 2, no. 1, pp. 97–106, 2016.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proc. Computer Vision and Pattern Recognition*, 2013, pp. 1139–1146.

[15] M. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *CVPR*, 2008.

[16] A. Patron, M. Marszalek, A. Zisserman, and I. Reid, "High five: Recognising human interactions in TV shows," in *Proc. BMVC*, 2010, pp. 50.1–11, doi:10.5244/C.24.50.

[17] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ser. ICPR '04, 2004, pp. 32–36.

[18] G. Yu, J. Yuan, and Z. Liu, "Propagative hough voting for human activity detection and recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 1, pp. 87–98, 2015.

[19] M. Raptis, I. Kokkinos, and S. Soatto, "Discovering discriminative action parts from mid-level video representations," in *CVPR*, 2012, pp. 1242–1249.

[20] S. Ma, J. Zhang, N. Ikizler-Cinbis, and S. Sclaroff, "Action recognition and localization by hierarchical space-time segments," in *ICCV*, 2013, pp. 2744–2751.

[21] A. Kläser, "Learning human actions in video," Ph.D. dissertation, Université de Grenoble, jul 2010.

[22] I. Everts, J. van Gemert, and T. Gevers, "Evaluation of color STIPs for human action recognition," in *CVPR*, 2013, pp. 2850–2857.

[23] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *CVPR*, 2011, pp. 3361–3368.

[24] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, pp. 60–79, 2013.

[25] S. Ma, L. Sigal, and S. Sclaroff, "Space-time tree ensemble for action recognition," in *CVPR*, 2015.

[26] G. Yu, J. Yuan, and Z. Liu, *Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ch. Propagative Hough Voting for Human Activity Recognition, pp. 693–706.

[27] A. Gaidon, Z. Harchaoui, and C. Schmid, "Activity representation with motion hierarchies," *International Journal of Computer Vision*, vol. 107, no. 3, pp. 219–238, 2014.