Evaluation of Local Spatio-temporal Salient Feature Detectors for Human Action Recognition

Amir H. Shabani[1,2] David A. Clausi[1] John S. Zelek[2]

Vision and Image Processing Lab.[1], Intelligent Systems Lab.[2]
University of Waterloo, ON, Canada N2L 3G1
{hshabani,dclausi,jzelek}@uwaterloo.ca

Abstract

Local spatio-temporal salient features are used for a sparse and compact representation of video contents in many computer vision tasks such as human action recognition. To localize these features (i.e., key point detection), existing methods perform either symmetric or asymmetric multi-resolution temporal filtering and use a structural or a motion saliency criteria. In a common discriminative framework for action classification, different saliency criteria of the structured-based detectors and different temporal filters of the motion-based detectors are compared. We have two main observations. (1) The motion-based detectors localize features which are more effective than those of structured-based detectors. (2) The salient motion features detected using an asymmetric temporal filtering perform better than all other sparse salient detectors and dense sampling. Based on these two observations, we recommend the use of asymmetric motion features for effective sparse video content representation and action recognition.

1 Introduction

Local spatio-temporal salient features have been widely used for sparse and compact representations of video content in many computer vision applications such as human action recognition [1, 2, 3, 4], video super-resolution [5], unusual event detection [6], human-computer interaction [7], and content-based video retrieval [8]. These features are typically localized in spatio-temporal key points where a sudden change in both space and time occurs. For example, 3D Harris corners occur when a spatially salient structure such as a corner changes its motion direction. This detector thus localizes the start/stop of local motion events in the video.

The salient features in a video represent the local video events which occur at different spatial and/or temporal scales. The spatial scale refers to the size of the body limbs or the subject as a whole which might vary across individuals and also by distance from the camera. The temporal scale refers to the fact that different people perform a given (sub-)action with different speed. In absence of any knowledge about these scales, a multi-scale analysis of the video signal is required to detect the features at different spatio-temporal scales. Effective feature detection is important for compact video content representation and consequently, action recognition, for example.

Detection of multi-scale salient features consists of three main steps: (1) spatio-temporal scale-space representation of the video signal, (2) saliency map construction, and (3) non-maxima suppression. Existing spatio-temporal feature detectors can be divided into two main categories: structured-based or motion-based. The structured-based feature detectors such as 3D Harris [9, 1] and 3D Hessian [3] are more selective of salient structures for which they incorporate different saliency criteria, but are limited to using just a symmetric 3D Gaussian filtering for a scale-space representation. The motion-based feature detectors such as Cuboids [4] and asymmetric motion features [2] localize the salient motion events in a video by treating the time domain different from space and hence, they are more consistent with human motion perception [10, 11, 12].

This paper evaluates the performance of both structured-based and motion-based feature detectors in a common framework for action recognition. To perform a fair comparison, we employ the standard discriminative bag-of-words action recognition framework [1] in which these features are utilized to learn the set of action prototypes and represent the action contents. Our objective is to find out which feature detector is the most effective method for video representation in an action classification application.

The rest of this paper is organized as follows. Section 2 reviews the existing salient spatio-temporal feature detec-

tors in video. Section 3 categorizes the existing detectors into structural-based and motion-based detectors. In this section, we briefly explain several examples from each category. Section 4 presents the human action recognition datasets, the evaluation framework, and the experimental results for performance evaluation of the different detectors. Finally, Section 5 summarizes the results.

2 Related work

Drawing inspiration from the usefulness of local multi-scale salient features for object recognition [13, 14], an immediate extension has been developed for spatio-temporal feature extraction for action recognition and for video analysis in general.

To extend 2D salient features to video, most of existing methods consider the sequence of images (2D+t) as a 3D object. As the 2D feature detectors select mainly salient structures in a still image, their extensions to 3D is considered as structured-based feature detectors. For example, Laptev et al. [9] extended the 2D Harris corner detector to 3D by performing 3D Gaussian filtering and computing the cornerness saliency criteria for a 3D autocorrelation matrix. In contrast, there are few methods that treat time domain different from space and detect motion-based salient features. For example, Dollar et al. [4] performed symmetric temporal Gabor filtering to detect salient motions referred to as Cuboids. Shabani et al. [12] used the difference of Poisson as the time-causal filter to detect opponent-based motion features. Recently, Shabani et al. [2] proposed a novel asymmetric multi-resolution temporal filtering to detect asymmetric motion features.

Our objective in this paper is to provide a fair and complete comparison of both structured-based and motion-based feature detectors for action classification. There are two very close publications to our evaluations in this paper. (1) Wang et al. [15] compared the performance of different sparse spatio-temporal key point detectors and dense sampling for human action classification. The dense samples detected at regular 3D points performed better on more realistic datasets such as UCF sports [16] and HOHA [17] which are collected from Youtube and Hollywood movies, respectively. However, it did not perform the best categorization of choreographed atomic actions in the KTH dataset [18]. The authors then concluded that the dense sampling method performs better on the real-world videos, but not on simple videos. (2) Shabani et al. [12, 2] evaluated the importance of temporal filtering in salient motion-based feature detection. They showed that the asymmetric temporal filters result in detection of motion features with higher precision rate and higher robustness under geometric changes such as camera view change or affine transformation [2]. Moreover, the asymmetric motion features [2] provide higher classification accuracy compared to features detected using a symmetric temporal filter such as Gabor (i.e, Cuboids [4]).

This paper can be considered as an extension to the existing spatio-temporal feature detectors evaluation papers. More specifically, in addition to motion-based features in [2], we also include the structured-based features in the comparison. Moreover, we set the number of spatio-temporal scales fixed for all the detectors. This is in contrast with the evaluation in [15] in which different feature types are detected at different set of spatio-temporal scales (e.g., Cuboids at one scale, 3D Harris at twelve scales) and as a result, the comparison is not consistent. Performing this complete evaluation determines the most effective feature detection method for action recognition. To this end, we use the standard discriminative bag-of-words recognition approach for the action classification. In this framework, the action primitives are learnt using the salient features of all the samples in the training set. An action is then represented globally as the frequency histogram of the appearance of the local features in the whole video.

3 Salient Feature Detectors

Existing spatio-temporal salient feature detectors can be categorized into two sets depending on whether detection of a salient structure is of interest or the detection of a salient motion is relevant. The differences come into the type of video filtering and the saliency criteria they use. The video filtering at different spatio-temporal scales provide a multi-resolution representation of the video contents from which features at different scales are detected. The saliency criteria determines which type of features will be chosen in their local spatio-temporal neighborhood. The salient features are localized at key points which are detected by performing non-maxima suppression in search window of $(3, 3, 3)$ [2].

In this section, we briefly explain different examples of both structured-based and motion-based feature detectors.

3.1 Structured-based features

To detect spatio-temporal structured-based features, existing methods treat the time domain as the third dimension of space and hence, they apply the same scale-space filter in space and time directions. That is, similar to the spatial Gaussian filtering, a temporal Gaussian is applied in the time direction [1, 3].

In this section, we briefly explain the extension of Harris corners and Hessian blobs to 3D which have been already used in action recognition literature [1, 3]. We also introduce the extension of 2D KLT (Kanade-Lucas-Tomasi) [19] features to 3D for action recognition.

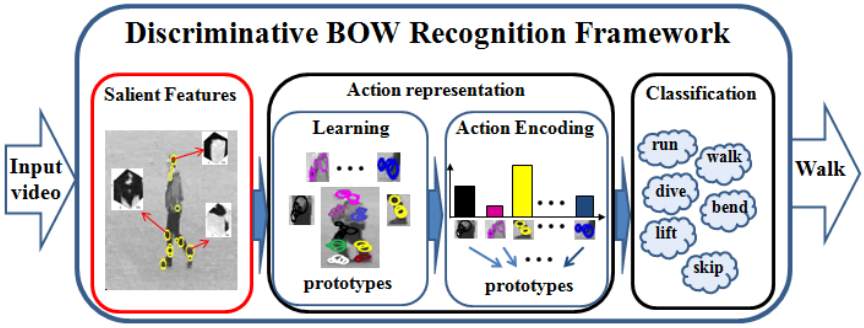


Figure 1. Standard discriminative bag-of-words framework for action recognition. The focus of this paper is to compare different salient feature detectors. These features encode the local video events and are used to learn the set of action prototypes (i.e., visual words or action primitives) during training. An action is represented by encoding its salient features over the prototypes. Finally, a classifier such as SVM determines the label of an unknown action.

3.1.1 3D Harris

Laptev et al. [9, 1] extended the Harris corner criteria from 2D image to 3D to extract corresponding points in a video sequence. To this end, the original video signal $I(x, y, t)$ is smoothed using a spatial Gaussian G_σ and a temporal Gaussian kernel G_τ using the convolution $L = G_\sigma * G_\tau * I$. The autocorrelation matrix $A = L_d^T \times L_d$ is then computed from the spatio-temporal derivative vector $L_d = [L_x, L_y, L_t]$. To compare each pixel to the neighborhood pixels a spatio-temporal Gaussian weighting $G_{2\sigma} * G_{2\tau}$ is then applied.

$$M = G_{2\sigma} * G_{2\tau} * A = G_{2\sigma} * G_{2\tau} * \begin{bmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_y L_x & L_y^2 & L_y L_t \\ L_t L_x & L_t L_y & L_t^2 \end{bmatrix} \quad (1)$$

The autocorrelation matrix M defines the second moment approximation for the local distribution of the gradients within a spatio-temporal neighborhood. Using the eigen-values $\lambda_1, \lambda_2, \lambda_3$ of the Harris matrix M, one can compute the spatio-temporal corner map C in which the corners are magnified and the rest are weakened ($k = 0.0005$).

$$C = \det(M) - k(trace^3(M)) = \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3 \quad (2)$$

3.1.2 3D Hessian

Willems et al. [3] extended 2D Hessian features to 3D by applying (an approximation) of 3D Gaussian filter and used the determinant of the Hessian matrix (3) as the saliency criteria. The points with high-value determinant ($S = \|det(H)\|$) represent the center of the ellipsoids (3D blob-like structures) in the video.

$$H = \begin{bmatrix} L_{xx} & L_{xy} & L_{xt} \\ L_{yx} & L_{yy} & L_{yt} \\ L_{tx} & L_{ty} & L_{tt} \end{bmatrix} \quad (3)$$

3.1.3 3D KLT

2D KLT features [19] have been widely used in many computer vision tasks such as tracking and structure from motion [20]. The 3D KLT [21] is the extension of its counterpart from 2D and it can be detected at multiple spatial and temporal scales. To this end, a family of scale-space representation of the video is obtained by performing 2D spatial Gaussian filtering G_σ and a temporal Gaussian filtering G_τ. At each scale, the 3D KLT saliency criteria is applied on the 3D autocorrelation matrix A to keep the points with the minimum of the eigen values above a threshold (i.e., $min(\lambda_1, \lambda_2, \lambda_3) > \alpha$). The 3D KLT features are then localized at points with maximum saliency value in their spatio-temporal neighborhood..

3.2 Motion-based features

The motion-based feature detectors perform the biologically-consistent Gaussian filtering for space, but they might use different temporal filters. The temporal filter might be symmetric or asymmetric. More specifically, consistency with the human's motion perception [10, 11] requires that the response to a periodic motion be mapped to a constant value. Moreover, the mapped representation should have the same value for two stimuli with different phases, but same motion patterns (i.e., it should be phase insensitive [11]). The filtering response should also be contrast-polarity insensitive [10, 11] to make sure that this representation is not sensitive to the polarity of the moving

stimuli versus background. For these phase and contrast-polarity insensitivity requirements, an energy model which induces quadrature-pair temporal filtering (i.e., two filters with 90 degree phase difference) is required [11]. The summation of the squared responses of the quadrature filters induces the energy map from which the salient motion features are detected.

In this section, we briefly explain different motion-based feature detectors which use a symmetric or an asymmetric temporal filter.

3.2.1 Cuboids (symmetric) motion features

Dollar et al. [4] used the energy field of temporal Gabor filtering on the spatially Gaussian smoothed video $R = (G_\sigma * F_{even} * I)^2 + (G_\sigma * F_{odd} * I)^2$ to extract the Cuboids centered at the spatio-temporal key points in the energy map. To detect the Cuboids at multiple scale, we performed the video filtering at different spatial (σ) and temporal (τ) scales which will be introduced in Section 4.2. Note that the even component (4) and odd component (5) of the complex Gabor filter are 90^o in phase difference and are essential to gain phase-insensitive motion map [12].

$$F_\tau^{even}(t) = cos(\omega_0 t) e^{-\frac{t^2}{2\tau^2}} \tag{4}$$

$$F_\tau^{odd}(t) = sin(\omega_0 t) e^{-\frac{t^2}{2\tau^2}} \tag{5}$$

3.2.2 Asymmetric motion features

Both Gaussian and Gabor are biologically consistent for spatial image filtering, but they are symmetric and non-causal which makes them not consistent with the temporal sensitivity of the human visual system [10, 11] and not feasible with the V1 cells physiology [22]. With this motivation from biological vision, Shabani et al. [23, 12, 2] advocate the use of time-causal video filtering for salient feature detection. Extending the spatial scale-space filtering to time, but with the time-causality constraint, the authors developed a new time-causal multi-resolution temporal filter based on the RC circuit theory [2]. The resulting filter is an asymmetric sinc filter $K(t;\tau)$ with a quadrature pair obtained form its convolution with the Hilbert transform (i.e., $K_h(t;\tau) = K(t;\tau) \star h(t)$ in which $h(t) = \frac{1}{\pi t}$ [10]).

$$K(t;\tau) = sinc(t-\tau)\, S(t) \tag{6}$$

where $S(t)$ denotes the Heaviside step function (i.e., $S(t) = 1,\ \forall t \geq 0$ and it is zero otherwise). Note that the shape of the asymmetric sinc kernel changes as a function of the temporal scale τ. More specifically, at finer scales, the kernel is more skewed towards the times before the peak of the kernel. The shape change of the asymmetric sinc filter with the scale increase results in detection of a wide range of motion features from asymmetric to symmetric motions. The performance comparison of the asymmetric sinc filtering with two other asymmetric filters of truncated exponential and Poisson [24] and with the symmetric Gabor [4] shows its higher efficiency [2]. In fact, the features detected using asymmetric sinc show higher precision rate, higher reproducibility under different geometric variations, and higher action classification rate [2]. From now on, we coin these features as asymmetric motion features.

We consider the performance comparison of three structured-based features (3D Harris, 3D Hessian, and 3D KLT) and two motion-based features (Cuboids and asymmetric motion features) for action recognition.

4 Experimental setup and results

This section presents our common action classification framework for feature evaluation, the data sets, and the action classification results using different salient features.

4.1 Action classification framework

For action classification, we incorporated the standard discriminative bag-of-words (BOW) setting [4, 15, 12] as shown in Fig. 1. In this framework, salient features are detected at multiple spatial and temporal scales (σ, τ) to localize the video events of different scales. Action prototypes/primitives are then learnt using the standard vector quantization procedure by clustering the features of all training samples into 1000 groups, experimentally [15, 2]. The clustering is performed by 10 times running the K-means algorithm with random seed initialization and keeping the result with the lowest error [15]. The clusters represent action primitives and are referred to as visual words in the BOW framework. Most existing methods select the number of cluster experimentally [15, 2], typically in order of 1000. To obtain a better statistics, we vary the number of clusters form 500 to 1500 with interval of 100 and report the average classification results.

By assigning each salient feature to its closest cluster (i.e., visual words), a global representation of an action is the frequency histogram of the appearance of the features in the whole video clip. The L_1 normalized frequency histogram is finally considered as the compact signature of the action. Finally, a nonlinear SVM (support vector machine) with the radial basis function (RBF) is utilized for the matching of action signatures (S_i, S_j). The parameter γ of the RBF (7) is learnt through cross validation using the LibSVM toolbox [25].

$$K_{RBF}(S_i, S_j) = e^{-\gamma |S_i - S_j|^2} \tag{7}$$

4.2 Spatio-temporal scales

To have a fair comparison, all the feature detectors use the same spatial ($\sigma_x = \sigma_y = \sigma_i$ per pixels) and temporal scales (τ_i per frames) in their scale-space video filtering. We consider combination of three spatial scales and three temporal scales, according to $2\ (\sqrt{2})^i, i = \{0, 1, 2\}$ formula [14, 2]. Note that the minimum spatial scale $\sigma_0 = 2$ pixels determines the maximum spatial frequency of $0.5\ cycles/pixel$(i.e., the highest spatial resolution is one cycle each two pixels). With 25 frames per second, the maximum temporal frequency of $12.5\ cycles/sec$ is obtained at $\tau = 2$. After video filtering at each spatio-temporal scale, the saliency map is computed using the corresponding saliency criteria of each feature detector (e.g., the cornerness (2) for the 3D Harris). To detect the key points and hence, localize the salient features, we perform non-maxima suppression in the spatio-temporal search window of $(3 \times 3 \times 3)$. To describe the motion and appearance of each detected feature, we use the 3D SIFT descriptor [26] which has shown promising performance in encoding the spatio-temporal histogram of oriented gradients in the feature's extension $(6\sigma \times 6\sigma \times 6\tau)$ [26].

4.3 Datasets

Three benchmark human action recognition datasets have been used for the performance evaluation of different detectors.

The **KTH data set** [18] consists of six actions (running, boxing, walking, jogging, hand waving, and hand clapping) with 600 choreographed video samples. Twenty-five different subjects perform each action in four different scenarios: indoors, outdoors, outdoors with scale change (fast zooming in/out) and outdoors with different clothes. According to the initial citation [18], the video samples are divided into a test set (9 subjects: 2,3,5,6,7,8,9,10, and 22) and a training set (the remaining 16 subjects).

The **UCF Sports dataset** [16] includes actions such as diving, golf swing, kicking, lifting, riding horse, run, skate boarding, swing baseball, and walk with 150 video samples collected from the Youtube website. This dataset is more challenging due to diverse ranges of views and scene changes with moving camera, clutter, and partial occlusion. A horizontally flipped version of each video is also used during training to increase the data samples [15]. Two version of this dataset has been used in the literature. The original authors [16] categorized this dataset into 9 classes, but recent publications [27, 28, 29, 15] split the samples of "swing" category into two categories of "swing on the pommel horse" and "swing at the high bar". We use leave-one-out (without considering the flipped samples for testing) protocol and report our results for both protocols.

The **Hollywood dataset** [30] consists of eight human actions (answer phone, get out the car, hand shake, hug a person, kiss, sit down, sit up, and stand up) from 32 Hollywood movies. The dataset is divided into a test set obtained from 20 movies and the (clean) training set obtained from 12 movies different from the test set. There are 219 sample videos in the training set and 211 samples in the test set.

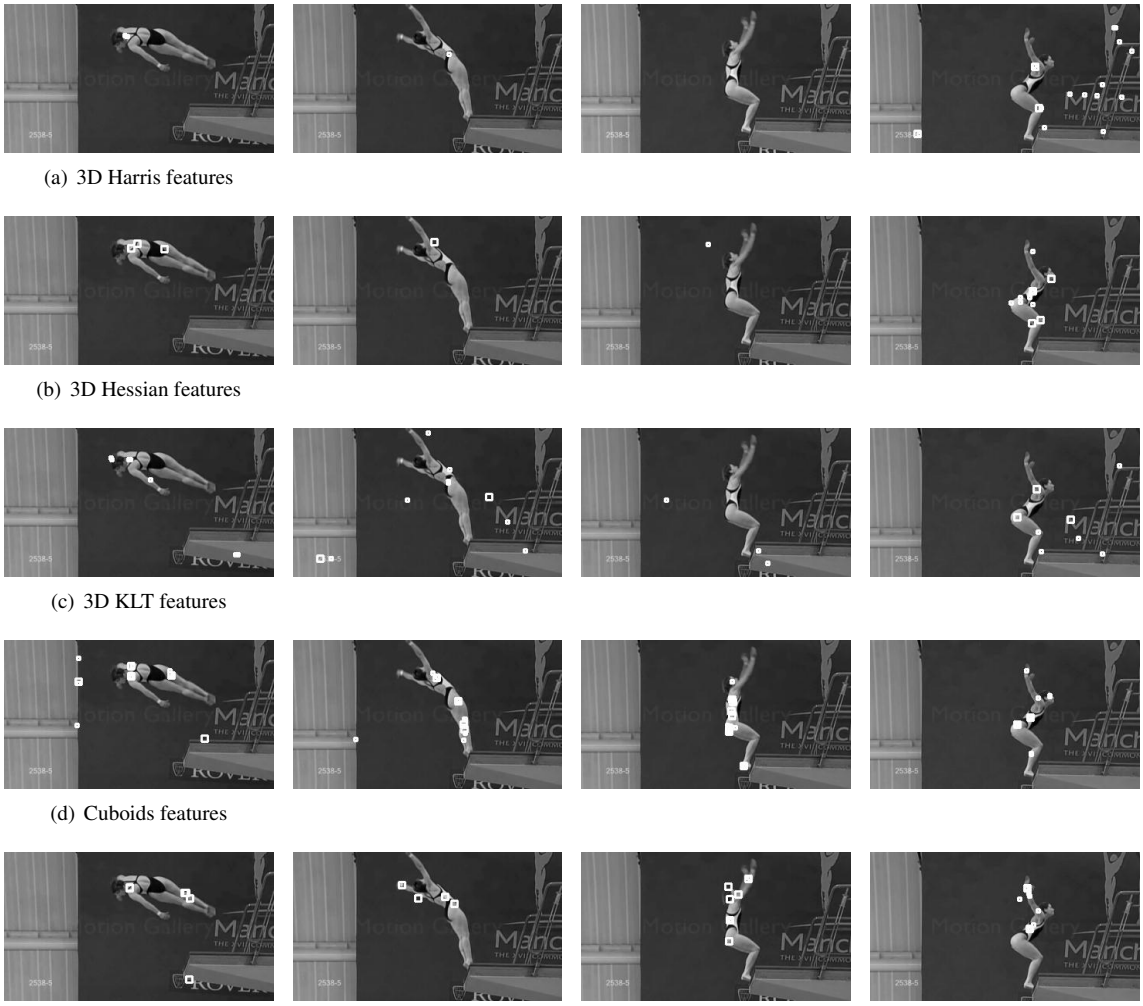
4.4 Action recognition results

Figure 2 shows the 2D projection of different spatio-temporal salient features on a sample frames from "diving" action from the UCF sports dataset [16]. In this video, the camera is following the athlete and that is a challenge for any local salient feature detector. In fact, the salient feature detector do not perform any background subtraction for video segmentation and hance, as a result, they find some features from the background (i.e., false positives). In this sample video, as it can be seen, among the structured-based features, the 3D Hessian has less false positive detection compared with the 3D Harris and the 3D KLT detector. In contrast, the motion-based detectors detect most of the features form the moving limbs. Among all the detectors, the asymmetric motion features are more localized on the moving limbs with much less false positives from the background. Note that one could perform camera motion compensation to reduce the false detection, but standard velocity-adaptation approaches such as Galilean transformation are computationally expensive and has not been successful in feature detection literature [1].

Table 1 presents the average classification accuracy of using different detectors on three different benchmark datasets. As can be seen, in all three datasets, the motion-based salient features perform better than the structured-base salient features. Among the structured-based salient features, the 3D Harris [9] and 3D KLT perform better than 3D Hessian [3]. Among the motion-based features, the asymmetric motion features [2] provide higher classification accuracy than the symmetric Cuboids [4]. These results support the importance of using motion-based features for video content representation and more specifically, the use of asymmetric temporal filtering to extract a wider range of motions from asymmetric to symmetric at different scales.

4.5 Comparison with other methods

Table 2 presents the classification rate of using asymmetric motion features and other published methods on three different datasets. As can be seen, the asymmetric features provide the highest accuracy on both the UCF and HOHA datasets. On the KTH dataset, our 93.7% accu-



(a) 3D Harris features

(b) 3D Hessian features

(c) 3D KLT features

(d) Cuboids features

(e) Asymmetric motion features

Figure 2. 2D projection of different multi-scale local salient features on sample frames of a "diving" action from the UCF sports dataset [16]. From top row to bottom row, the features are (a) 3D Harris, (b) 3D Hessian, (c) 3D KLT, (d) Cuboids, and (e) asymmetric motion features. Among all the detectors, the asymmetric motion features are more localized on the moving body limbs with much less false positives from the background.

racy is comparable with 94.2% [29] accuracy obtained using joint dense trajectories and dense sampling which require much more computation time and memory compared to our sparse features. In a comparable setting with [15], the asymmetric motion features perform better than other salient features and dense sampling.

5 Conclusion

In a common discriminative framework for action classification, we compared different salient structured-based and motion-based feature detectors. For performance evaluation, we used three benchmark human action recognition datasets of the KTH, UCF sports, and Hollywood. In all three datasets, the motion-based features provide higher classification accuracy than the structured-based features. More specifically, among all of these sparse feature detectors, the asymmetric motion features perform the best as they capture a wide range of motions from asymmetric to symmetric. With much less computation time and memory usage, these sparse features provide higher classification accuracy than the dense sampling as well. Based on our experimental results, we recommend the use of asymmetric motion filtering for effective salient feature detection, sparse

Table 1. Average classification accuracy on different datasets using the features detected by different methods. The accuracy variation is in order of 0.01 and is not reported here. Note that motion-based detectors perform better than structured-based detectors on all three datasets. Moreover, the asymmetric features provide higher classification accuracy than the symmetric Cuboids features.

Dataset		Structured-based fetaures			Motion-based features	
		3D Harris	3D Hessian	3D KLT	Cuboids	Asymmetric
KTH [18]		63.5 %	67.5 %	68.2 %	89.5 %	**93.7** %
UCF sports [16]	*9 classes*	72.8 %	70.6 %	72.6 %	73.3 %	**91.7** %
	10 classes	73.9 %	70.2 %	72.5 %	76.7 %	**92.3** %
HOHA [30]		58.1 %	57.3 %	58.9 %	60.5 %	**62** %

Table 2. Comparison of different published methods for the human action classification on the KTH [18], the UCF sports [16], and the HOHA [30] datasets. In this table, the bold italic items show the original protocol of the dataset introduced by their corresponding authors.

Method	*KTH*	*UCF sports*		*HOHA*
		9 classes	10 classes	
Laptev et al. [30]	-	-	-	38.4 %
Schuldt et al. [18]	71.7 %	-	-	-
Rodriguez et al. [16]	86.7 %	69.2 %	-	-
Shabani et al. [2]	93.3 %	91.5 %	-	-
Wang et al. [29]	**94.2** %	-	88.2 %	-
Wang et al. [15]	92.1 %	-	85.60 %	-
Willems et al. [3]	88.3 %	-	85.60 %	-
Asymmetric motions	93.7 %	**91.7** %	**92.3** %	**62** %

video content representation, and consequently, action classification.

Acknowledgment

The authors would like to thank both GEOIDE (Geomatics for Informed Decisions), supported by the Natural Science and Engineering Research Council (NSERC) of Canada, and the Ontario Centres of Excellence (OCE) for financial support of this project.

References

[1] I. Laptev, B. Caputo, C. Schuldt, and T. Lindeberg. Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding*, pages 207–229, 2007.

[2] A. H. Shabani, D. A. Clausi, and J. S. Zelek. Improved spatio-temporal salient feature detection for action recognition. *British Machine Vision Conference, Dundee, UK*, Sep. 2011.

[3] G. Willems, T. Tuytelaars, , and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. *European Conference on Computer Vision, Marseille, France*, pages 650–663, Oct. 2008.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal filters. *IEEE International Workshop VS-PETS, Beijing, China*, pages 65–72, Aug. 2005.

[5] O. Shahar, A. Faktor, and M. Irani. Space-time super-resolution from a single video. *IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO*, pages 3353–3360, June 2011.

[6] T. Zhang, S. Liu, C. Xu, and H. Lu. Boosted multi-class semi-supervised learning for human action recognition. *Pattern Recognition*, 44:23342342, 2011.

[7] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28:976–990, 2010.

[8] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank. A survey on visual content-based video indexing and re-

trieval. *IEEE Transactions on Systems, Man, and Cybernetics*, 41(6):797–819, 2011.

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

[10] A. B. Watson and A. J. Ahumada. Model of human visual-motion sensing. *Journal of Optical Society of America*, 2(2):322–342, 1985.

[11] E. H. Adelson and J.R. Bergen. Spatio-temporal energy models for the perception of motion. *Optical Society of America*, pages 284–299, 1985.

[12] A. H. Shabani, J.S. Zelek, and D.A. Clausi. Human action recognition using salient opponent-based motion features. *IEEE Canadian Conference on Computer and Robot Vision, Ottawa, Canada*, pages 362 – 369, May 2010.

[13] Harris C. and M.J. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, pages 147–152, 1988.

[14] D. G. Lowe. Distinctive image features from scale-invariant key points. *International Journal of Computer Vision*, 60:91–110, 2004.

[15] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. *British Machine Vision Conference, London, UK*, Sep. 2009.

[16] M.D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. *IEEE Conference on Computer Vision and Pattern Recognition, Alaska*, pages 1–8, June 2008.

[17] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. *IEEE Conference on Computer Vision and Pattern Recognition, Miami, Florida*, pages 2929–2936, June 2009.

[18] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. *IEEE International Conference on Pattern Recognition, Cambridge, UK*, 3:32–36, Aug. 2004.

[19] J. Shi and C. Tomasi. Good features to track. *IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA*, pages 593–600, 1994.

[20] N. Govender. Evaluation of feature detection algorithms for structure from motion. *3rd Robotics and Mechatronics Symposium (ROBMECH), Pretoria, South Africa*, page 4, Nov. 2009.

[21] Y. Kubota, K. Aoki, H. Nagahashi, and S.I. Minohara. Pulmonary motion tracking from *4D-CT* images using a *3D-KLT* tracker. *IEEE Nuclear Science Symposium Conference Record,Orlando, FL*, pages 3475–3479, Oct. 2009.

[22] B.M. ter Haar Romeny, L.M.J. Florack, and M. Nielsen. Scale-time kernels and models. *Scale-Space and Morphology in Computer Vision, Vancouver, Canada*, pages 255–263, July 2001.

[23] H. Shabani, D.A. Clausi, and J.S. Zelek. Towards a robust spatio-temporal interest point detection for human action recognition. *IEEE Canadian Conference on Computer and Robot Vision, Kelowna, BC*, pages 237–243, May 2009.

[24] T. Lindeberg and D. Fagerstrom. Scale-space with causal time direction. *European Conference on Computer Vision, Cambridge, UK*, pages 229–240, April 1996.

[25] C.C. Chang and C.J. Lin. *LIBSVM: a library for support vector machines*, 2001. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[26] P. Scovanner, S. Ali, and M. Shah. A 3-*D*imensional *SIFT* descriptor and its application to action recognition. *ACM Multimedia, Augsburg, Germany*, pages 357–360, Sep. 2007. Code available at `http://www.cs.ucf.edu/~pscovann/`.

[27] A. Gaidon, Z. Harchaoui, and C. Schmid. A time series kernel for action recognition. *British Machine Vision Conference, Dundee, UK*, Sep. 2011.

[28] 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. *IEEE Conference on Computer Vision and Pattern Recognition, Colorado Spring*, pages 3361–3368, June 2011.

[29] H. Wang, A. Klaser, C. Schmid, and C.L. Liu. Action recognition by dense trajectories. *IEEE Conference on Computer Vision and Pattern Recognition, Colorado Spring*, pages 3169–3176, June 2011.

[30] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska*, pages 1–8, 2008.