Systematic Evaluation of Spatio-temporal Features on Comparative Video Challenges

Julian Stöttinger[1,3], Bogdan Tudor Goras[2], Thomas Pöntiz[3],
Allan Hanbury[4], Nicu Sebe[5] and Theo Gevers[6]

[1]CVL, Institute for Computer-Aided automation, TU Vienna,
[2]Faculty of Electronics, Telecommuniction and Informatics, Tech. University of Iasi,
[3]CogVis Ltd., Vienna, [4]IR Facility, Vienna,
[5]Dept. of Information Eng. and Computer Science, University of Trento,
[6]Faculty of Science, University of Amsterdam

Abstract. In the last decade, we observed a great interest in evaluation of local visual features in the domain of images. The aim is to provide researchers guidance when selecting the best approaches for new applications and data-sets. Most of the state-of-the-art features have been extended to the temporal domain to allow for video retrieval and categorization using similar techniques to those used for images. However, there is no comprehensive evaluation of these. We provide the first comparative evaluation based on isolated and well defined alterations of video data. We select the three most promising approaches, namely the Harris3D, Hessian3D, and Gabor detectors and the HOG/HOF, SURF3D, and HOG3D descriptors. For the evaluation of the detectors, we measure their repeatability on the challenges treating the videos as 3D volumes. To evaluate the robustness of spatio-temporal descriptors, we propose a principled classification pipeline where the increasingly altered videos build a set of queries. This allows for an in-depth analysis of local detectors and descriptors and their combinations.

1 Introduction

The bag-of-words approach, has been successfully adapted to the use of visual vocabularies describing images [1]. One central question for this approach is the choice of the right visual features. For set of local features the aim is to describe visual data successfully in a discriminative and robust way. Additionally, the data to be processed should be reduced as much as possible and should lead to a robust representation of the video. Video features based on local 3D patches are a popular representation for videos in tasks in retrieval, recognition and categorization (e.g. [2–5]). The most promising approaches for spatio-temporal features are corner detectors [6], blob detectors [7], periodic spatio-temporal features [8], volumetric features [9], and spatio-temporal regions of high entropy [10].

Recent work [11] points out that throughout the literature many experiments are not comparable. As such, the justification of specific properties of detectors and descriptors advocated in the literature is often insufficient. For example, results are frequently presented for different data-sets such as the KTH data-set [8, 12, 13, 4, 5, 7, 14], the Weizmann data-set [15] or the aerobic actions data-set [10]. Nevertheless, in that evaluation

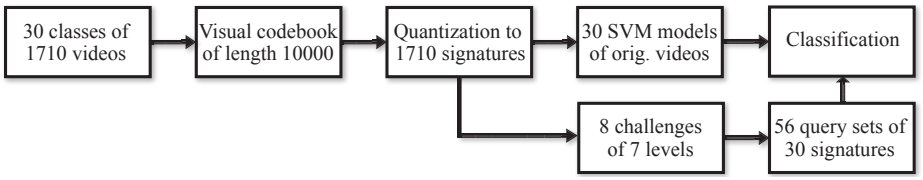

Fig. 1. Experimental setup to test the description's robustness against visual alterations.

paper, combinations of detectors and descriptors are only measured on their final classification accuracy on the mentioned data-sets. A principled evaluation of every step of a matching framework, as is successfully done in "2D" images (e.g. [16]), is missing for "3D" video matching so far.

Therefore, we propose a new way for the evaluation of video retrieval approaches: We divide the evaluation of detectors and descriptors into two independent tasks. For detection, we use a repeatability measurement in 3D similar to [7]. For the descriptions we propose a pipeline to identify the robustness of local spatio-temporal descriptions in a principled way. These two tasks are measured by their performance under alterations of the visual input data. Therefore, we use a publicly available dedicated on-line data-set[1] providing 30 classes of videos [17]. Every video undergoes 8 types of transformations denoted as *challenges*. Each challenge is applied at 7 levels of increasing impact on the video leading to 1710 videos in total (compare Fig. 1). We use the original videos as ground-truth and observe to what extent the features change under the challenges. Example frames can be found in Fig. 2.

(a) HD movie (b) surveillance video (c) TV show

Fig. 2. Example videos and example transformations.

We follow [11] and use the best performing approaches Harris3D, Hessian3D and the Gabor detector and HOG/HOF, SURF3D (also referred to as *extended SURF*) and HOG3D for our evaluation on videos. We use the same parameters and the same implementations.

[1] www.feeval.org

The paper is organized as follows. The chosen features are described in more detail in Section 2. The experimental setup is described in Section 3. Results are given in Section 4. Section 5 gives a critical discussion and conclusions.

2 Spatio-temporal Features

An extension of the Harris corner detector [18] is the **Harris3D** detector [6]. The authors compute a spatio-temporal second-moment structure tensor at each video point using independent spatial and temporal scale values σ, τ, a separable Gaussian smoothing function G, and space-time gradients L. Extending the scale space to the temporal domain, we add the temporal variance τ^2 to get $L_{\mathbf{x},\sigma^2,\tau^2} = G_{\mathbf{x},\sigma^2,\tau^2} * f^t_{\mathbf{x}}$ and use the image data of the corresponding video frame f^t. The spatio-temporal Gaussian kernel is defined as

$$G_{\mathbf{x},\sigma^2,\tau^2} = \frac{1}{2\pi\sigma^4\tau^2} e^{-\frac{x^2+y^2}{2\sigma^2} - \frac{t^2}{2\tau^2}} \tag{1}$$

It is separable and thus can be calculated for each dimension on its own and in parallel. This extension gives then the structure tensor M for every location and scale. The final locations are extracted by applying $H = \det(M) - k \cdot \mathrm{trace}^2(M)$ and extracting the positive maxima of the corner function H. Points are extracted at multiple scales based on a regular sampling of the scale parameters s, t as suggested by the authors. We use the original implementation[3] and its settings $k = 0.0005$, $s^2 = 4, 8, 16, 32, 64, 128$, $t^2 = 2, 4$ with a detection threshold of 10^{-9}.

The **Hessian3D** detector [7] is the spatio-temporal extension of the Hessian blob detector [19]. The saliency of a location is given by the determinant of the 3D Hessian matrix. For efficiency, box-filter operations are applied on an integral video structure on multiple scales. Each octave is divided into 5 scales, with a ratio between subsequent scales in the range [1.2; 1.5] for the inner 3 scales. A non-maximum suppression algorithm selects the common extrema over space, time and scales: (x,y, t,s, τ). It is defined by the structure tensor Γ

$$\Gamma = \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix} \tag{2}$$

where the *strength* S of an interest point is given by its tensor determinant $S = |det(\Gamma)|$. We use the authors' implementation[4] with the suggested parameters.

The **Gabor** detector is a set of spatial Gaussian convolutions and temporal Gabor filters [8]. The Gabor filters give a local measurement focusing not only on local changes in the temporal domain, but prioritize repeated events of a fixed frequency. The function gives $R_{\mathbf{x}t\sigma\tau\omega} = (f^t_{\mathbf{x}\sigma} * G_{\mathbf{x}\sigma} * H^{ev}_{t\tau\omega})^2 + (f^t_{\mathbf{x}\sigma} * G_{\mathbf{x}\sigma} * H^{od}_{t\tau\omega})^2$ where the 2D Gaussian smoothing is only applied in the spatial domain, whereas the two filters H^{ev} and H^{ov} are applied in the temporal domain only. H^{ev} and H^{ov} are the quadrature pair of 1D Gabor filters. The set of functions is available on-line as a toolbox[2]. As suggested and used in previous evaluations, we chose $\sigma = 3$ and $\tau = 4$.

[2] `vision.ucsd.edu/~pdollar/toolbox/doc/index.html`

To describe the detected patches by local motion and appearance, [4] compute histograms of spatial gradients and optical flow accumulated in space-time neighborhoods of detected interest points referred to as **HOG/HOF**. HOG results in a descriptor of length 72, HOF in a descriptor of length 90. For proper performance they are simply concatenated. The descriptor size is defined by $D_x(\sigma) = D_y(\sigma) = 18\sigma$, $D_t(\tau) = 8\tau$. The approach is inspired by the SIFT descriptor. In the experiment, the grid parameters $n_x, n_y = 3$, $n_t = 2$ as suggested in [4]. The binaries are available online[3].

Willems et al. [7] proposed the **SURF3D** (ESURF) descriptor which extends the image SURF descriptor to videos. An image patch is represented by a 288 dimensional vector of weighted sums of uniformly sampled responses of Haar-wavelets. The binaries are also available[4]. 3D patches are divided into $n_x \times n_y \times n_t$ cells. The size of the 3D patch is given by $D_x(\sigma) = D_y(\sigma) = 3\sigma$, $D_t(\tau) = 3\tau$. For the feature descriptor, each cell is represented by a vector of weighted sums $v = (\sum d_x, \sum d_y, \sum d_t)$ of uniformly sampled responses of the Haar-wavelets d_x, d_y, $d_t$along the three axes.

For the third descriptor in the evaluation we use the **HOG3D** [13]. This is based on histograms of 3D gradient orientations efficiently computed using an integral video representation. It leads to a descriptor of length 960.

3 Experimental Setup

In this section, the experimental set-up used throughout the evaluation is described . Section 3.1 presents an overview of the evaluation data-set used. In Section 3.2, the methodology for the detector evaluation is given. The pipeline and the parameters of the classification task for the descriptor evaluation is described in detail in Section 3.3.

3.1 Video Data-set and Features

Our experiments aim to quantify the robustness of the state-of-the-art spatio-temporal features described in the previous section. We challenge the robustness of these approaches on the FeEval data-set [17][1], which consists of 1710 videos of about 20 seconds each. Starting with 30 short clips from HDTV shows, Hollywood movies of a *full HD* resolution of 1920×1080, and surveillance videos, the full FeEval dataset is created as follows: (1) Every video undergoes 8 types of systematic alterations denoted as challenges. The challenges are noise, increasing lightness, decreasing lightness (darkness), median filtering, compression, scale and rotation, and reduction in frames per second. (2) Each challenge is applied at 7 levels of increasing impact, and encoded by a parameter (see Fig. 2). The parameters and the challenge abbreviations used throughout the experiments are given in Tbl. 1. This leads to about 34 Gigabytes (GB) of H.264 compressed video material.

[3] www.irisa.fr/vista/Equipe/People/Laptev/download.html#stip
[4] homes.psat.kuleuven.be/~gwillems/research/Hes-STIP/

Transformation	Abbreviation	Range
Gaussian blur: σ in pixels	blur	3 - 21
H.264 compression	compr	60 - 0
Noise in %	noise	5 - 35
Median Filter: σ in pixels	median	2 - 8
Increasing lightness in %	lighten	+30% - +90%
Decreasing lightness in %	darken	-30% - -90%
Frames per Second	fps	20 - 3
Scale + Rotation in degrees	scalerot	90% & 10° - 30% & 70°

Table 1. Video transformations for each of the 30 videos.

3.2 Detector Evaluation

To evaluate the robustness of the three detectors Harris3D, Hessian3D, and Gabor, we measure their robustness or *repeatability* for each altered video with respect to its corresponding original video. Each of the 30 original video is regarded as a *boolean* 3D volume V_{o_i}, $i = \overline{1, 30}$, sized according to the frame resolution and the total number of frames. $V_{o_i} = 1$ if a voxel is being detected by a feature or 0 otherwise. Every of the m detected feature $\xi_{c,1..m}$ in an altered video is defining a cuboid in space. Per repeatability test, we map the cuboid $\xi_{c,j}$ to V_o to get its position and expanion in the original video's volume V_o denoted as $\xi'_{c,j}$. This is done by applying its homography matrix Ω to $V_o \leftarrow \Omega * V_c$ For the challenge of scale and rotation, we use the provided "2D" matrices defined by the parameters given in Tbl. 1, as the alteration is per frame only and does not affect the temporal configuration. For the challenge of decreasing frames per second, we regard it as a simple scaling in the temporal direction and apply it on the t expansion of V_c only. Overlap ϱ of feature j is then defined by

$$\varrho = \frac{V_o \cap \xi'_{c,j}}{\upsilon(\xi_{t,i})} \tag{3}$$

where $\upsilon(\xi_{t,i})$ is the volume of the transformed feature's cuboid. The final repeatability score of a video is defined by the number of matched features divided by the total number of features in the challenge video.

3.3 Descriptor Evaluation

We want to test the ability of state-of-the art spatio-temporal descriptors to what extent they maintain their robustness under alteration of their input videos. We aim to test their performance in a large scale video classification experiment where the training data consists of 30 original videos forming 30 classes of challenges. For the three descriptors HOG/HOF, SURF3D and HOG3D and the combination with the detectors we carry out the following set-up:

We form a visual codebook of 10000 words by clustering all the features of the data-set with the *kshift* [20][5] algorithm. In contrast to many other clustering implementations, the data-set can be larger than the memory. For every cluster center, it is only necessary to have the *next* feature in the memory, not the whole data-set. It is feasible to cluster 45 GB of 960 dimensional features within 20 hours using 2 X5560@2.8GHz processors (4 cores each). A video's signature is built by quantizing its features to the

[5] `www.cogvis.at`

codebook by the cluster center with the nearest Euclidean distance. For the training set, we use the 30 original videos with their normalized signatures of a length of 10000 each as ground truth classes. For every class, we train a linear one-against-all SVM model equally weighting every class. For this setup, the model is similar to a nearest neighbor classification. We are using the well known LibSVM library[6] with default parameters. For the 8 challenges with 7 levels, we build 56 test sets of equal size to be evaluated. The experimental question is then until which alteration the description is still able to discriminate against the other videos and under which circumstances it fails. When an altered video is successfully classified as its original video, the description is regarded as robust to the alteration. In this context, the classification performance according to the alterations gives then the descriptor robustness in the challenge.

4 Results

Starting with the repeatability experiments in the following Section 4.1 we are able to evaluate the robustness of the detections of state-of-the-art spatio-temporal features. In Section 4.2 the three descriptors are evaluated in a classification experiment.

4.1 Detector Evaluation

Regarding the overall repeatability performance the Hessian3D detector outperforms the Harris3D detector, whereas the Gabor detector shows to be significantly less robust. The mean results on varying ϱ are given in Fig. 3. The single-scale Gabor detector is not much affected by the change of the overlap criterium, as the large number of small features tends to be matched almost perfectly or not at all. This is of course different for the multi-scale approaches Harris3D and Hessian3D, where different sizes of features are matched.

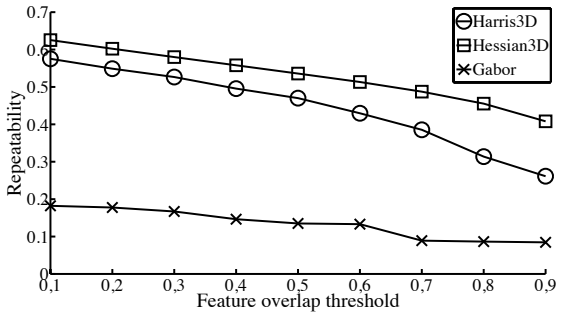

Fig. 3. Mean repeatability results for the whole data-set over varying overlap ϱ.

Hessian3D has the best mean repeatability and performs best throughout the experiments. However, it provides a richer representation as its coverage is almost 10

[6] www.csie.ntu.edu.tw/~cjlin/libsvm

times larger than Harris3D, thus making the probability for a geometrical match higher. Still, Harris3D performs comparably similar, which coincides closely to the evaluation of their 2D counterparts in [16] As we observe in Fig. 4(a), Harris3D and Hessian3D are almost equally robust to increasing blur. This also holds for increasing compression shown in Fig. 4(b). The two detectors are very robust to increasing compression, showing similar results on 2D images [16]. This is an important observation, since the spatio-temporal structure tensor has more degrees of freedom and a much bigger dataset than it has been done for 2D repeatability. In contrast to 2D detectors, the Harris3D and Hessian3D show to be very sensitive to change of lightness (see Fig. 4(e) and 4(f)). The number of features decreases rapidly with the decrease of contrast. This is the only challenge where the Gabor detector outperforms the other approaches in robustness at level 7. The decrease of frames per second (see Fig. 4(g)) can be seen as scaling in the temporal domain. As the approaches are not scale invariant, they perform worse than their 2D counterparts. Hessian3D regarding the most scales of the approaches evaluated remains rather stable until level 3, which is the reduction from 25fps to 13fps. Therefore the standard sampling rate of 2 for the Hessian3D approach can be easily set to 4 without a significant loss in performance, disregarding 50% of the data right away. For scale and rotation, Gabor and Harris perform poorly compared to the Hessian3D which is able to maintain a repeatability rate of 0,41 for a video scaled by a factor of 0.3 and rotated by 70 degrees. Harris3d and Gabor are very sensitive to noise, Hessian3D remains stable showing a repeatability of 0,62 with 35% of noise in the video. For increasing median filtering, Harris3D is equally robust as the Hessian3D.

Following these results, the following for noisy video data is proposed: Gaussian blur degrades the detections severely therefore it should not be used in pre-processing videos. Hessian3D on noise performs more robust than on blurred data. Gabor detections are neither reliable on noisy or blurred data. When using the Harris3D detector, it is recommended to use the median filter to remove the noise in advance.

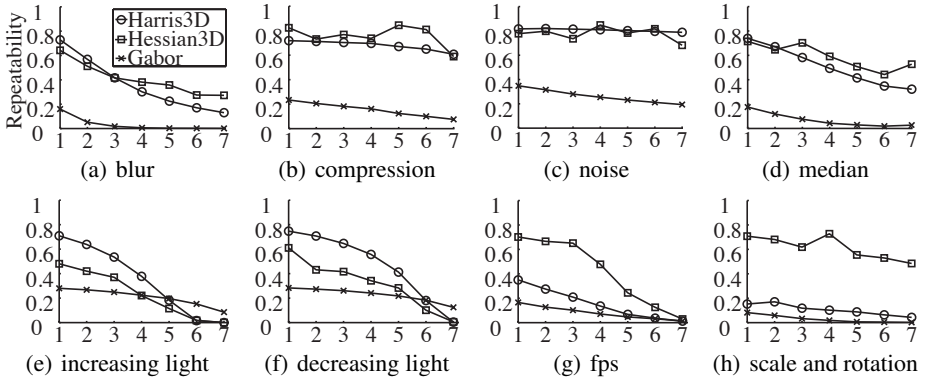

Fig. 4. Mean repeatability ($\varrho = 0.6$) of 30 videos per challenge. Legend is found in (a).

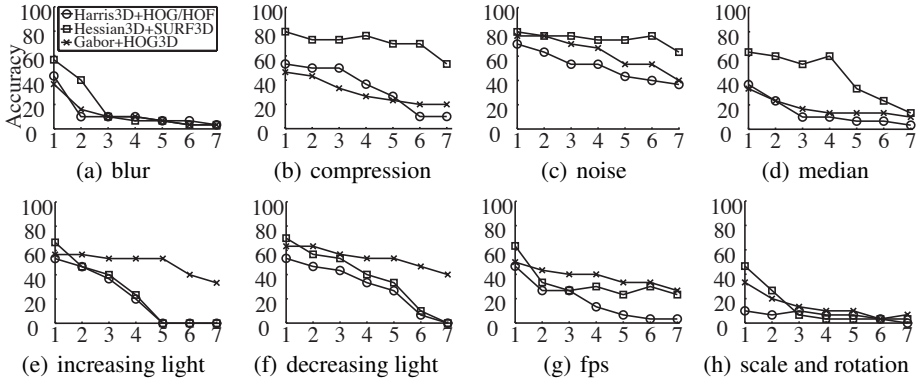

Fig. 5. Classification accuracy with increasing alterations of the query images with suggested descriptor and detector combinations. Legend is found in (a).

4.2 Descriptor Evaluation

Summary results are shown in Tbl. 2 Results per challenge are shown in Fig. 5. In Fig. 6 results of the experiments using the HOG3D descriptor are given. The combination of Harris3D and HOG3D outperforms other approaches.

	Classification accuracy			**Mean precision**			**Mean recall**		
	Harris3D	Hessian3D	Gabor	Harris3D	Hessian3D	Gabor	Harris3D	Hessian3D	Gabor
HOG/HOF	23,57	-	-	19,40	-	-	23,57	-	-
SURF3D	-	39,52	-	-	40,46	-	-	44,80	-
HOG3D	**49,76**	37,96	34,75	**42,40**	38,80	28,15	**49,76**	42,20	35,30

Table 2. Overview experimental results descriptor evaluation.

As already argued in the previous section, Gaussian blur decreases the representation of the videos significantly. As seen in Fig. 5(a), the classification accuracy goes towards the prior probability of 3%. This is different for the HOG3D descriptor. For all detectors, there is a significant gain in classification performance, especially for the Harris3D+HOG3D raising to a mean accuracy of 54,76%.

Similar behavior is observed for change of lightness: For HOG/HOF and SURF3D, the classification accuracy goes down rapidly, whereas the HOG3D descriptor provides a stable description on data of varying contrast. Gabor+HOG3D outperforms these approaches (see Fig. 5(e) and 6(e)). When combining the detectors with HOG3D, we observe a correlation with the repeatability experiments of changing lightness. With a more stable descriptor, the more repeatable representation influences the classification performance. This does not hold for the fps challenge (see Fig. 5(g) and 6(g)). There is no correlation between detector robustness and classification performance. This suggests that none of the descriptors is scale invariant to a satisfying extent.We deduce that for performance reasons, detectors can be applied on a reduced data-set but the local description has to be performed on full resolution.

Descriptors revealed to be more robust to increasing noise than the local detectors. Worst performing Harris3D+HOG/HOF reaches a mean accuracy of 51,43%. Hessian3D + SURF3D remains almost stable throughout the challenge (see Fig. 5(c)).

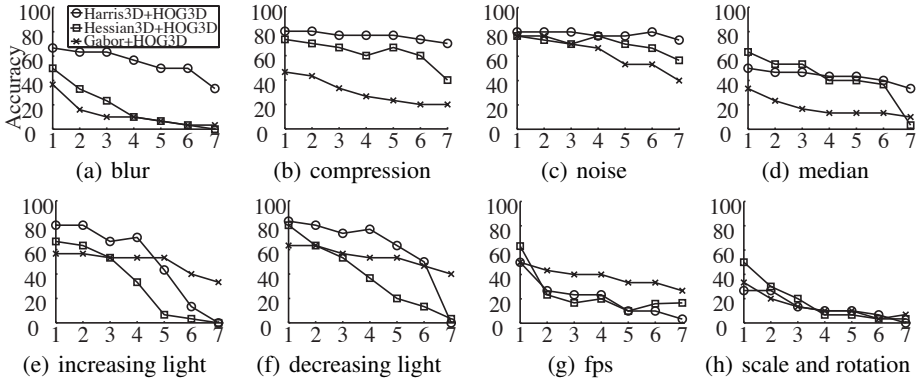

Fig. 6. Classification accuracy with increasing alterations of the query images with detectors and HOG3D descriptor. Legend is found in (a).

HOG3D shows to be more robust than HOG/HOF (see Fig. 6(c)), but decreases the performance for the Hessian3D. It is shown that SURF3D is more robust to noise than HOG3D in this context. Regarding noise reduction using the median filter (see Fig. 5(d) and 6(d)) performance decreases more than for the noise challenge. HOG/HOF and HOG3d are sensitive to the filtering, SURF3D performs best coherent to the repeatability rate of its detector. Increasing compression does not affect the description performance of the HOG3D and the SURF3D descriptor. Even strong JPEG artifacts are described in a stable and discriminative way (see Fig. 5(b) and 6(b)). For level 7 of the challenge, the data is compressed up to 10% of the original file size.

To sum up the evaluation, we interpret the results categorizing them to simple votes according to the challenges. '-' denotes sensitivity, '+' robustness to the challenge. '+/-' refers to undecided decision or room for improvements in the algorithmic details of the approach. Our final suggestions are given in Tbl. 4.2.

	Detector Robustness			**Descriptor Robustness**		
	Harris3D	Hessian3D	Gabor	HOG/HOF	SURF3D	HOG3D
Gaussian blur	+/-	+/-	-	-	-	+/-
H.264 compression	+	+	-	-	+	+
Noise	-	+	-	+/-	+	+
Median Filter	+	+	-	-	+/-	+/-
Increasing lightness	+/-	+/-	+/-	-	-	+
Decreasing lightness	+/-	+/-	+/-	-	-	+
Frames per Second	-	+	-	+/-	+/-	+/-
Scale & Rotation	-	+	-	+/-	+/-	+/-

Table 3. Final suggestions based on the evaluation.

5 Conclusion

In this work, we perform the first principled evaluation of spatio-temporal features using comparative challenges inspired by prior evaluation of local 2D image features. For

detector robustness, we experienced comparable results for spatio-temporal features with their image counterparts. Generally, it showed to be worse to reduce noise in input data than to let the features take care of it on their own. For change of lightness, both the Harris3D and the Hessian3D are more sensitive than their 2D counterparts. Description is most stable using the HOG3D descriptor, outperformed by the SURF3D descriptor in the challenges of compression, noise and median filtering. The high dimensionality of the HOG3D descriptor of 960 compared to 288 of the SURF3D descriptor is a drawback in terms of the complexity of all succeeding operations and should be considered when choosing the most appropriate descriptor.

References

1. Cula, O.G., Dana, K.J.: Compact representation of bidirectional texture functions. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on **1** (2001) 1041
2. Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J.: Automatic annotation of human actions in video. In: ICCV. (2009)
3. Junejo, I., Dexter, E., Laptev, I., Pérez, P.: View-independent action recognition from temporal self-similarities. PAMI (2009)
4. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008) 1–8
5. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR. (2004)
6. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV. (2003)
7. Willems, G., Tuytelaars, T., Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: ECCV. (2008) 650–663
8. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS. (2005) 65–72
9. Ke, Q., Kanade, T.: Quasiconvex optimization for robust geometric reconstruction. In: ICCV. (2005) 986 – 993
10. Oikonomopoulos, A., Patras, I., Pantic, M.: Kernel-based recognition of human actions using spatiotemporal salient points. In: CVPR. (2006) 151
11. Wang, H., Ullah, M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC. (2009)
12. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: ICCV. (2007) 1–8
13. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC. (2008) 995–1004
14. Wong, S.F., Cipolla, R.: Extracting spatiotemporal interest points using global information. In: ICCV. (2007) 1–8
15. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. PAMI **29** (2007) 2247–2253
16. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. IJCV **65** (2005) 43–72
17. Stöttinger, J., Zambanini, S., Khan, R., Hanbury, A.: Feeval - a dataset for evaluation of spatio-temporal local features. In: ICPR. (2010)
18. Harris, C., Stephens, M.: A combined corner and edge detection. In: AVC. (1988) 147–151
19. Lindeberg, T.: Feature detection with automatic scale selection. IJCV **30** (1998) 79–116
20. Pönitz, T., Donner, R., Stöttinger, J., Hanbury, A.: Efficient and distinct large scale bags of words. In: AAPR. (2010)