FeEval - A Dataset for Evaluation of Spatio-temporal Local Features

Julian Stöttinger, Sebastian Zambanini, Rehanullah Khan
Institute of Computer-Aided Automation
Vienna University of Technology

Allan Hanbury
Information Retrieval Facility, Vienna

Abstract

The most successful approaches to video understanding and video matching use local spatio-temporal features as a sparse representation for video content. Until now, no principled evaluation of these features has been done. We present FeEval, a dataset for the evaluation of such features. For the first time, this dataset allows for a systematic measurement of the stability and the invariance of local features in videos. FeEval consists of 30 original videos from a great variety of different sources, including HDTV shows, 1080p HD movies and surveillance cameras. The videos are iteratively varied by increasing blur, noise, increasing or decreasing light, median filter, compression quality, scale and rotation leading to a total of 1710 video clips. Homography matrices are provided for geometric transformations. The surveillance videos are taken from 4 different angles in a calibrated environment. Similar to prior work on 2D images, this leads to a repeatability and matching measurement in videos for spatio-temporal features estimating the overlap of features under increasing changes in the data.

1. Introduction

Video understanding gains great attention in current computer vision research. With growing on-line data sources of videos, big digital private video archives and the need for storage and retrieval of surveillance videos, automated video understanding becomes necessary.

Techniques such as the bags-of-words approach are originally inspired by text retrieval. These have been extended to "2D" techniques on images. These approaches are now successfully carried out in both the spatial and the temporal domains for action recognition, video understanding and video matching (e.g. [2, 5, 9, 14]). Common in these works is the first step of the approach where a set of local features is extracted.

Recent work [15] evaluates spatio-temporal features on their matching performance on different datasets. They state that in the literature many experiments are not comparable as they differ in their experimental settings and classification techniques. However, they do not evaluate the robustness of the features themselves, but only in the context of the final classification accuracy of a complex experiment. Although we know from image retrieval that the choice of features has a significant impact on the overall result of the bags-of-words approach, classification accuracy is a strong hint, not an in-depth analysis of the quality of a representation.

We propose a way to evaluate the quality of these features in an independent way from the framework or application. It provides the first database to evaluate extracted features for their stability and invariance in a spatio-temporal context, called *FeEval*. Every transformation denotes one *challenge* and is well defined. For the geometric cases all homography matrices are known. The change of noise, light, compression or frames per second are applied reproducibly according to the parameters given. The dataset consists of 30 original videos. Per video, 8 transformations are applied in 7 increasing steps, leading to a total of 1710 videos. FeEval is available on-line[1].

The paper is organized as follows. The following section gives an overview of current research in spatio-temporal features and how the proposed dataset contributes to the state of the art. Section 3 gives an overview of existing datasets. Section 4 defines the naming conventions and the applied transformations. Section 5 concludes.

[1] www.feeval.org

2. Evaluation of Spatio-temporal Features

The most promising approaches for spatio-temporal features are spatio-temporal corners [8], periodic spatio-temporal features [1], volumetric features [6] and spatio-temporal regions of high entropy [12]. Following [11], we desire a stable representation which is invariant to lighting conditions, view point, quality of encoding, resolution and frames per second. However, until now there is no principled evaluation of the robustness of spatio-temporal features available: Evaluation is done by measuring the overall performance of the application itself [15]. An evaluation of a complex framework only by its final performance does not give full insight into the performance of the chosen features. Subsequent operations (clustering, classification) are arbitrarily chosen and use empirically found parameters. Moreover, experiments in the literature are carried out with different classification algorithms tainting the experimental insights.

Recent work [15] is concerned with the progress in the community of action recognition. They state that while specific properties of detectors and descriptors have been advocated in the literature, their justification is often insufficient. Limited and non-comparable experimental evaluations are used in current papers. For example, results are frequently presented for different datasets such as the KTH dataset [1, 4, 7, 9, 14, 16, 17], the Weizmann dataset [3] or the aerobic actions dataset [12][2]. However, many results are incommensurable as they differ in their experimental setups. A principled evaluation of every step of a matching framework as it is successfully done in "2D" images (e.g. [11]) is missing for "3D" video matching.

[15] improve this fact and evaluate different combinations of spatio-temporal features, dense sampling and descriptors. Evaluation is done by their recognition performance for a bags-of-words classification by a χ^2-kernel SVM. The evaluation shows clearly that the right features have a significant impact on the matching performance. Nevertheless, some questions remain: Most of the parameters of the classification systems are chosen because of empirical estimation. In addition the clustering for the codebook generation is not consistent with prior work making it difficult to draw conclusions. Furthermore, most of the previous evaluations are reported for actions in controlled environments such as in the KTH and Weizmann datasets. It is therefore unclear how these methods generalize to action recognition in realistic setups [9, 13].

[2] www.eecs.berkeley.edu/Research/Projects/CS/vision/action/

3. Video Datasets

In this section we present a brief overview of existing action recognition datasets.

The **KTH actions dataset** [14][3] provides videos of six human action classes: walking, jogging, running, boxing, waving, and clapping. Each action class is performed repeatedly by 25 persons. The sequences were recorded in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. All the 2391 black and white sequences were taken over homogeneous backgrounds with a static camera with 25fps frame rate at a resolution of 160x120 pixels.

The **Weizmann dataset** [3][4] provides 90 videos of nine people at a resolution of 180x144 pixels at 50 fps. Action classes are run, walk, skip, jumping jack, jump forward on two legs, jump in place on two legs, gallop sideways, wave to hands and wave one hand. Perfect retrieval results are obtained from various authors.

The **UCF sport actions dataset** [13][5] contains ten different types of human actions with a great intra-class variety: swinging (on bar, pommel horse, floor), golf swinging, walking, diving, weight-lifting, horse-riding, running, skateboarding and kicking. It provides 182 video sequences at a resolution of 720x480 pixels

The **Hollywood2 actions dataset** [10][6] has been collected from 69 different Hollywood movies. There are 12 action classes: answering the phone, driving car, eating, fighting, getting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up. There are 69 movies divided into a training set (33 movies) and a test set (36 movies) resulting in a total of 3669 sequences. Train and test sets are obtained from a non overlapping set of Hollywood movies.

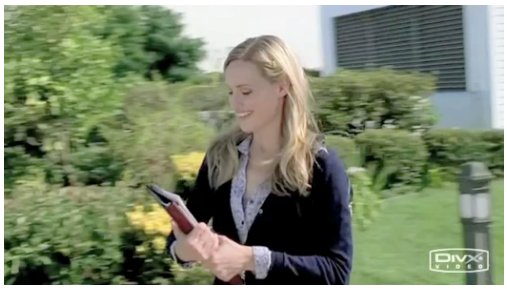


Figure 1. Video1, 624x352 HDTV show.

[3] www.nada.kth.se/cvap/actions/

[4] www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

[5] www.cs.ucf.edu/vision/public_html/

[6] pascal.inrialpes.fr/hollywood2/

Transformation	Abbreviation	1	2	3	4	5	6	7
Gauss σ in pixel	blur	3	6	9	12	15	18	21
Noise in %	noise	5	10	15	20	25	30	35
Darken: Lightness in %	dark	-30	-40	-50	-60	-70	-80	-90
Lighten: Lightness in %	light	30	40	50	60	70	80	90
Median Filter σ in pixel	median	2	3	4	5	6	7	8
H.264 quality	comp	60	50	40	30	20	10	0
Scale + Rotation in degree	scalerot	90% + 10°	80% + 20°	70% + 30°	60% + 40°	50% + 50°	40% + 60°	30% + 70°
Frames per Second	fps	20	15	13	10	7	5	3

Table 1. Video transformations for each of the 30 videos. Filename convention: "[number of video]-abbreviation-[number of column].mov"

4. FeEval

FeEval consists of 30 videos from HD TV shows, 1080p HD movies and surveillance videos. Every video undergoes 8 transformations with successive impact, denoted as challenges. This leads to a dataset of 1710 videos each of about 20 seconds. All videos are encoded with the H.264 codec and stored in a .mov Quicktime container with 25 fps.

10 videos are taken from 2 long running TV shows. Some challenges are visualized in Fig. 4. Using TV show material has several advantages: We are able to access a vast amount of video content of a manageable group of people (the TV show cast) over the time of several years. Additionally, the actors also appear in other shows and movies, making large scale person detection and recognition experiments possible.

Surveillance videos show 3 different people in a calibrated environment. The persons enter the lab, fall on the floor, get up and leave the scene again. Every fall sequence is taken from 4 different viewpoints (see Fig. 2). A 3D calibration target placed in the scene was used to achieve an accurate camera calibration with an average reprojection error of ~ 0.2 pixels. The calibration makes it possible to map world coordinates to respective image coordinates and consequently to recover 3D structure from the 2D images, which finally enables repeatability and robustness measurements among different viewpoints. All 4 camera projection matrices are available on the website.

The 1080p HD movies are challenging because of their high resolution of 1920x1080 pixels and therefore the high demand of memory and processing power. An example is given in Fig. 3. Run-time and scale invariance of spatio-temporal features can be evaluated on the state-of-the-art of the home entertainment formats.

Every challenge consists of 7 levels. An overview is given in Table 1. Additional annotation, the persons and actions in the videos are provided on the webpage.

The **Gaussian blur** challenge applies increasing Gaussian blur per color channel. The kernel size is increased by 3 pixels at every level, beginning with a size of 3 pixels leading to 21 pixels for the 7th level.

(a) 11.mov (b) 12.mov (c) 13.mov (d) 14.mov

Figure 2. Calibrated scene from 4 view points.

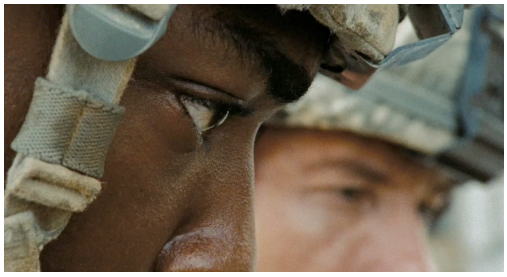


Figure 3. Video 27, 1080p HD movie.

Noise adds random values to the video. Beginning with 5% noise in every frame, the challenge increases the amount of noise for every step by 5% up to 35%.

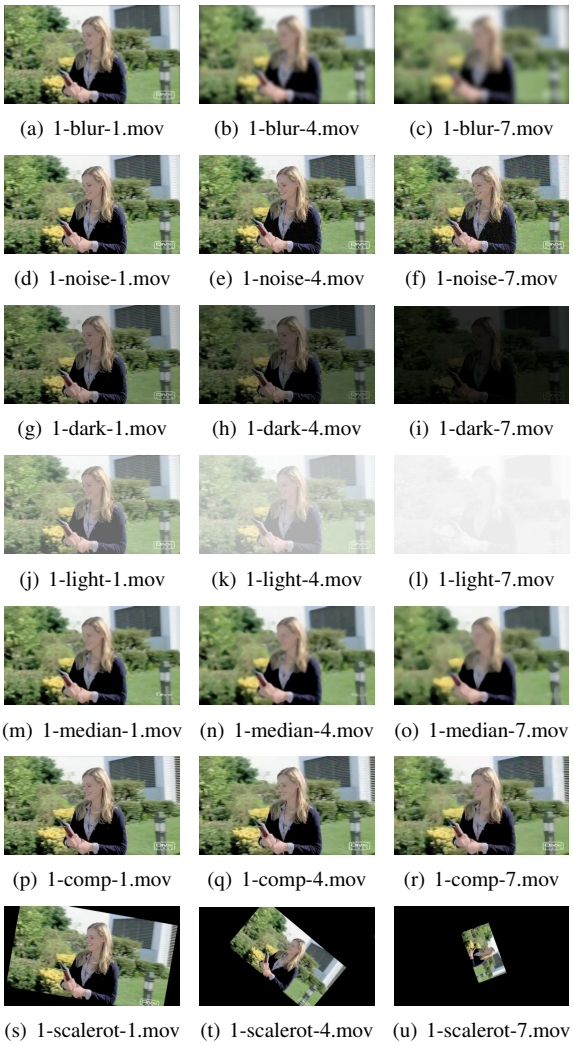
Change of lighting We darken and lighten the videos by changing the saturation of the colors to simulate increasing and decreasing lighting conditions. The change of lighting is applied from $\pm$ 30% to $\pm$ 90%.

The **median filter** is used to reduce speckle noise and salt and pepper noise effectively. We apply the filter with a kernel size from 2 pixels to 8 pixels.

To test the effect of **increasing compression**, we decrease the H.264 quality from 60 to 0 leading to a video with the occurence with strong JPEG artifacts and many wrong colors and edges.

Invariance to **scale and rotation** is evaluated by increasingly shrink the videos to a final size of 30% of the original size and rotate them by 10% for every level. The homography matrices are straightforward to estimate and given at the webpage.

To decrease the demand for storage space, surveillance videos are often handled with very few **frames per second**. For the challenge, the original 24 frames per second are reduced up to 3 frames per second.



(a) 1-blur-1.mov (b) 1-blur-4.mov (c) 1-blur-7.mov
(d) 1-noise-1.mov (e) 1-noise-4.mov (f) 1-noise-7.mov
(g) 1-dark-1.mov (h) 1-dark-4.mov (i) 1-dark-7.mov
(j) 1-light-1.mov (k) 1-light-4.mov (l) 1-light-7.mov
(m) 1-median-1.mov (n) 1-median-4.mov (o) 1-median-7.mov
(p) 1-comp-1.mov (q) 1-comp-4.mov (r) 1-comp-7.mov
(s) 1-scalerot-1.mov (t) 1-scalerot-4.mov (u) 1-scalerot-7.mov

Figure 4. Overview dataset of video 1.

5. Conclusion

We present a dataset to evaluate the robustness and invariance of spatial features against 8 challenges. For the first time, data for the evaluation of spatio-temporal features is available. For geometric transformations, homography matrices are provided. Furthermore, the videos have overlapping cast making it possible to evaluate action and person recognition.

In contrast to existing datasets, FeEval consists of videos of varying sources from surveillance cameras to high resolution movies. All the videos are in color and display a grand variety of persons, surroundings and lighting conditions. With this dataset of 1710 videos, we allow for a principled evaluation on generalized data by measuring the geometric repeatability and the description robustness against well defined challenges.

Acknowledgments

This work was partly supported by the Austrian Research Promotion Agency (FFG) project OMOR 815994, MuBisA 819862 and the CogVis[7] Ltd.

References

[1] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, pages 65–72, 2005.

[2] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *ICCV*, pages 1–8, 2009.

[3] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *PAMI*, 29(12):2247–2253, 2007.

[4] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, pages 1–8, 2007.

[5] I. Junejo, E. Dexter, I. Laptev, and P. Pérez. View-independent action recognition from temporal self-similarities. *PAMI*, 2009.

[6] Q. Ke and T. Kanade. Quasiconvex optimization for robust geometric reconstruction. In *ICCV*, pages 986 – 993, 2005.

[7] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, pages 995–1004, 2008.

[8] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, pages 432–439, 2003.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, pages 1–8, 2008.

[10] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, pages 2929–2936, 2009.

[11] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.

[12] A. Oikonomopoulos, I. Patras, and M. Pantic. Kernel-based recognition of human actions using spatiotemporal salient points. In *CVPR*, page 151, 2006.

[13] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. *CVPR*, 2008.

[14] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004.

[15] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[16] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, pages 650–663, 2008.

[17] S. F. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. In *ICCV*, pages 1–8, 2007.

[7] www.cogvis.at/