Compression-Based Clustering of Video Human Activity Using an ASCII Encoding

Guillermo Sarasa[1(✉)], Aaron Montero[1], Ana Granados[2], and Francisco B. Rodriguez[1]

[1] Grupo de Neurocomputación Biológica, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid, Spain
guillermo.sarasa@predoc.uam.es, montaaron@gmail.com, f.rodriguez@uam.es
[2] CES Felipe II, Universidad Complutense de Madrid, Aranjuez, Madrid, Spain
ana.granados@ajz.ucm.es
http://arantxa.ii.uam.es/∼gnb/

Abstract. Human Activity Recognition (HAR) from videos is an important area of computer vision research with several applications. There are a wide number of methods to classify video human activities, not without certain disadvantages such as computational cost, dataset specificity or low resistance to noise, among others. In this paper, we propose the use of the Normalized Compression Distance (NCD), as a complementary approach to identify video-based HAR. We have developed a novel ASCII video data format, as a suitable format to apply the NCD in video. For our experiments, we have used the *Activities of Daily Living Dataset*, to discriminate several human activities performed by different subjects. The experimental results presented in this paper show that the NCD can be used as an alternative to classical analysis of video HAR.

Keywords: Data mining · Normalized Compression Distance
Clustering · Dendrogram · Image processing
Human Activity Recognition · Silhouette Coefficient · Similarity

1 Introduction

Human Activity Recognition (HAR) [4,6,31] from videos represent a relevant area of computer vision research. Its utility in many areas has increased the demand of broader analysis in the field, producing an increase of publications related with Computer Vision in HAR [4,26,31]. Some of its applications are: human health care [17], video labeling [27,28], surveillance [21,26]and human-computer interaction [1,24], among others. There are many approaches in the literature to identify human activities from video with remarkable results. However, dealing with video implies solving certain issues that eventually lead to some drawbacks in the final systems of HAR video processing. Some examples are high computational costs, dataset specificity or the dependency of the temporal movement sequence.

V. Kůrková et al. (Eds.): ICANN 2018, LNCS 11140, pp. 66–75, 2018.
https://doi.org/10.1007/978-3-030-01421-6_7

Vision-based HAR can be summarized as a combination of extracting some features from a sequence, and discriminating between activities by means of a classification system. The most important difficulties of feature extraction in video processing are: (i) overlap and variability between and within classes, (ii) temporal differences between samples (iii) impact and complexity of the environment and (iv) quality of the data. As an example of the first problem, a video may contain activities that include similar movements (e.g. reading and using a tablet) but also can include activities that are carried out differently by different people (e.g. cooking). Following this last case we can find others examples of the second problem. Among others, the duration, repetition or even order of execution of an activity can differ greatly, causing variations in the temporal structure, or sequence, of the activity. Finally, the capability to identify the background depends on many factors such as color difference, movement of the camera, or even quality of the recorded video.

There are a considerable variety of methods that aim to solve these problems in the literature [4,6,31]. However, as we introduced before, the inherent drawbacks of these methods require additional adjustments in order to be used in a real-world application. In this work, we aim to use compression algorithms as a parameter free dissimilarity approach (among other reasons, see Sect. 2.1) to identify human activities in video files. The idea behind using a parameter free method is to identify the relevant information without performing any low level analysis on the data. This is to increase the applicability of the method (due to the lack of specificity and parameters) while decreasing its computational costs (that some times make the system prohibitive to real-world implementations). Also, the use of compression distances over video data represents a novel application with remarkable applications for video analysis.

In this work, we have developed a video-to-ASCII processing method to locate and convert the activity of the video files into suitable objects for a compression algorithm. In order to test the capabilities of these methodology, we have performed experiments over the *Activities of Daily Living Dataset* [22] (see Sect. 3). This dataset is composed of different videos of human activities, performed by different subjects. Each video is recorded from a fixed point of view and stored in *Audio Video Interleave* (AVI) format, using the Motion JPEG video codec. In our experiments we try to discriminate between each pair of activities, parsing each video into our ASCII video format and using a widely used compression distance (the so-called Normalized Compression Distance or NCD) together with a hierarchical clustering. The results obtained using our methodology report a good separability between most of the pairs of activities. These results suggest that this measure could be used as an alternative methodology to identify video HAR.

2 Methodology

As mentioned before, we have used the *Activities of Daily Living Dataset* [22]. This data set has been used in several studies on human activity recognition in

the literature [2,20]. In this Section we will introduce the compression distances, (as the methodology that we have used in this work) the methodology to convert video streams into ASCII objects and the clustering procedure to measure the identification capabilities of the NCD.

Fig. 1. Video activities examples, obtained from the *Activities of Daily Living Dataset* [22]. The five upper pictures belong to the activities labeled as: "answer phone", "chop banana", "dial phone", "drink water" and "eat banana". The five lower pictures belong to the activities labeled as: "eat snack", "look up in phonebook", "peel banana", "use silverware" and "write on whiteboard".

2.1 Normalize Compression Distances

Compression distances are dissimilarity measures that make use of compression algorithms to identify common properties between objects. These measures search for the information shared between files, and use it, to define how different, in general terms, two objects are. The Normalized Compression Distance (NCD), is a generalization defined in [8,19] that defines the distance between two objects x and y, as the relation between the size of each object compressed alone ($C(x)$ and $C(y)$), and the size of their concatenation (xy) compressed ($C(xy)$). Hence, if the concatenation of two objects can be compressed better than each object alone, it means that the objects share some information. The mathematical formulation of the NCD can be defined as:

$$NCD(x,y) = \frac{\max\{C(xy) - C(x), C(yx) - C(y)\}}{\max\{C(x), C(y)\}} \ ,$$

where C is a compression algorithm and $C(x)$ and $C(xy)$ are the size of the C-compressed versions of x and the concatenation of x and y, respectively. The NCD has been used in different areas of knowledge, with remarkable results, due to its high noise tolerance, wide applicability and capabilities among different types of data (audio, images, text, etc.). Among many others, compression distances have been used from document clustering [13,14] to spyware and phishing detection [7,18], image analysis [10,11,16,29],earth observation [5,15] and music clustering [12,25].

Due to the fact that compression distances are based on the skill of a compressor to identify similar features in big amounts of data, one would expect that video data should not be an exception. However, the video codecs used to store

video streams (sequence of images) in video files, already compress the information. In contrast to a text book or a bitmap picture, where the information is fully accessible, a video file contains the information compressed, making its identification by a compression algorithm almost impossible. The way in which the information is compressed, depends on the codec used for the video file. In the data used in this paper, each video sequence is stored using the Motion JPEG codec (one of the few lossless video codecs), which compress each frame individually as a separate image. This however is not the only issue that the NCD has with video objects. Among others, the high percentage of noise or the big heterogeneity of sizes, are examples of other drawbacks to applying NCD directly to the video format. For all these reasons, we propose a novel video ASCII representation, in order to mitigate some of these drawbacks.

2.2 Data Format: From Video to ASCII

In order to transform the activity videos into a format that could be appropriate to be used by compression algorithms, we have developed a video preprocessing method. The aim of this process is to extract the optical flow [3] of the video objects and to obtain the motion signature of each task that takes place in them. This motion signature is the one that will be encoded in ASCII format to be analyzed by the compressor. This encoding allows reducing the size of the original video files from 14.4–211 MBs to a fixed 17 KB for the ASCII format (which also solves the size problem mentioned before).

The video preprocessing consists of the following steps:

1. We extract 10 video frames from the video, equally separated in time, on which we perform a grayscale conversion, see panel (a) in Fig. 2
2. We calculate the optical flow (through Horn–Schunck method [30]) of the selected frames and apply a thresholding to obtain the image points with greater activity, see panel (b) in Fig. 2.
3. We divide the image into binary boxes (1 = movement, 0 otherwise) and calculate the total activity produced in each one of them. This will generate an activity map, see panel (c) of Fig. 2. The dimensions of the boxes used are 16×16 pixels.
4. We obtain the motion signature adding the different activity maps into a unique one, see panel (d) in Fig. 2.
5. We assign identifiers to each of the image boxes using a diagonal zigzag order (used in image encoding such as MPEG [23]), see panel (e) in Fig. 2.
6. Once the boxes are organized by means of the identifiers, we sort them according to the total activity (given by the optical flow) of each of the boxes. This is the information that will be stored into an ASCII file and, later on, analyzed by the NCD, see panel (f) in Fig. 2.

2.3 Clustering of ASCII Objects Using String Compression

Once the video objects have been parsed into our proposed ASCII video objects, it is necessary to define a methodology to measure the effect of the NCD into

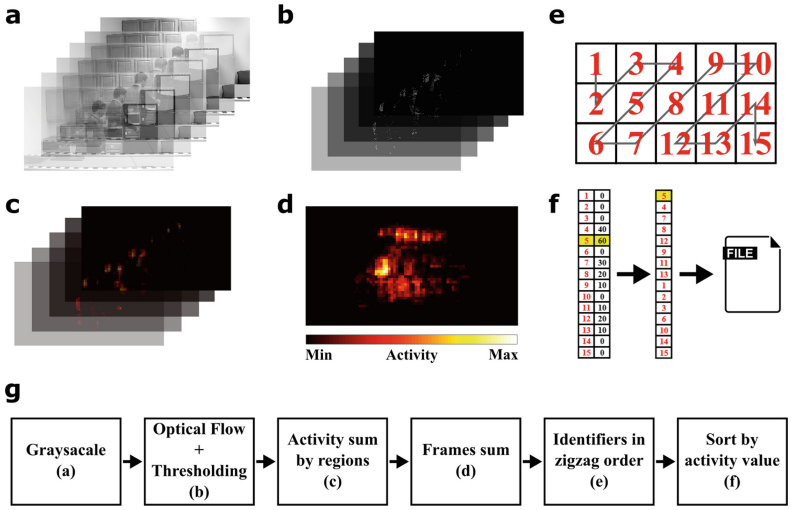

Fig. 2. Video preprocessing for conversion from the original AVI to ASCII format. Firstly, we extract a certain number of video frames and convert them to grayscale, panel (a). Secondly, we calculate the optical flow of these frames and apply a threshold on them to obtain the image points with greater activity, panel (b). Thirdly, we divide the image into boxes and calculate the total activity produced in each of them in order to generate an activity map, panel (c). Subsequently, we make the sum of the different activity maps to obtain the motion signature, panel (d). Finally, we read the image matrix (motion signature) in zigzag order, panel (e), and sort the information as a function of the total activity of each of the boxes, panel (f).

these new ASCII objects. Due to the NCD only reports a distance between two objects, we make use of a hierarchical clustering algorithm (based on the MQTC algorithm [8] from the CompLearn toolkit [9]) to parse the NCDs between objects into a dendrogram. For instance, given the case of a set of ASCII video objects, for two of the classes of Fig. 1, we can measure the NCDs between every pair of files and transform it into a hierarchical dendrogram. Finally, in order to measure how well each class is separated, we have made use of the Silhouette Coefficient (SC) (detailed in [14]) as an unbiased clustering quality measure.

3 Experiments and Results

For our experiments we have taken all the data provided by the Activities of Daily Living Dataset [22] to measure the capabilities of our methodology. This dataset includes 10 different tasks performed by 5 different subjects, 3 times each one of them. The objective in these experiments is to discriminate two sets of 15 objects each, from two classes of the videos of Fig. 1. In this figure, we show a representative frame of each class along with the names of the different tasks to classify. As an example to motivate the complexity of this problem, in Fig. 3, we

show 10 samples of processed activity maps before the zig-zag sort (described in Sect. 2.2) and the dendrogram produced by the NCD-driven clustering over the 30 video objects (15 of each class). The left figures are obtained from videos of two activities performed by 5 different subjects. In this figure, one can see that the activity classes have different signatures, but are not easily differentiable at simple sight. In order to identify these signatures we made use of a NCD-driven clustering (described in Sect. 2.3) which, as the right dendrogram of the figure shows, identify the two classes perfectly.

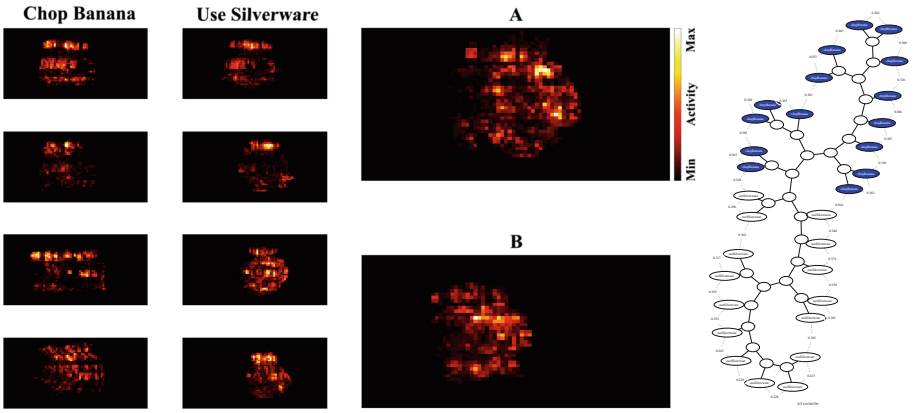

Fig. 3. Sample maps of activity for *Chop banana* and *Use silverware*, for different subjects. Each heatmap is produced by the process described in Sect. 2.2 until the zig-zag sort. This is equivalent to the *d* panel of Fig. 2. The right heatmaps, A and B, belong to *Use silverware* and *Chop banana*, respectively. As we can see, the classes are not easily differentiable at simple sight. The dendrogram of the figure shows how well our method identify each activity for all the subjects samples. The Silhouette Coefficient in this case is 0.51

In Fig. 4 one can see that the proposed format, together with the NCD, report remarkable task identification results for the majority of tasks pairs. However, there is some tasks that are more difficult to identify than others. For example, while "chopBanana" and "eatSnack" are very well separated, "peelBanana" and "eatSnack" are not. Following the first case ("chopBanana" and "eatSnack"), in Fig. 5 we show the dendrogram corresponding to the field marked with an X of Fig. 4, with and without our video-to-ASCII process (right and left dendrograms, respectively). One can see that the clustering is only achieved in the right dendrogram, where all the video objects are processed into the activity ASCII objects. Thus, the conversion of the video objects prove to be essential to the analysis.

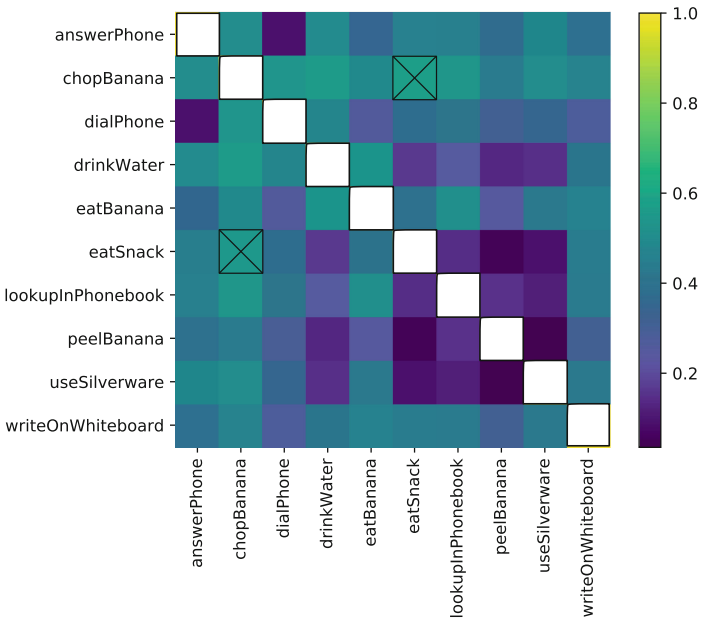

Fig. 4. Color map comparison of the clustering quality obtained from the different experiments. Each point of the map, corresponds to the S.C. obtained from parse the video to our video format (described in Sect. 2.2) and applying a NCD-driven clustering (described in Sect. 2.1). The diagonal of the matrix is not defined due to the fact that a task cannot be compared with itself. The dendrogram of the fields marked with an X is depicted in Fig. 5 right panel.

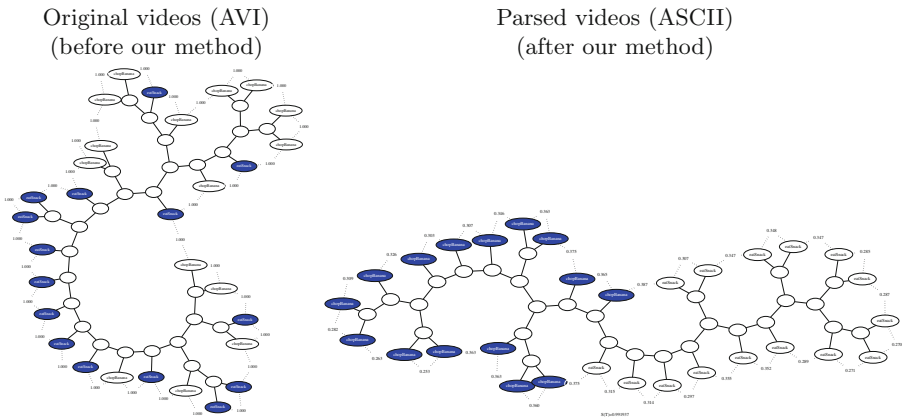

Fig. 5. Sample dendrograms, produced by the clustering of the activities: *Chop banana* and *Eat snack*, for the original files and the processed files. One can easily see that both activities are well separated in the right dendrogram (where the videos are transformed into our proposed format) while the left dendrogram (obtained from the original videos) reports almost no separability. Additionally, the Silhouette Coefficient for these dendrograms is 0.546 and 0.123, respectively. The right dendrogram corresponds to the fields marked with an X of Fig. 4.

4 Conclusions

The approach presented in this work aims to identify different human activities from video sequences addressing some of the drawbacks that classical systems have. The way in which we have performed that consist of adapting a generic, low costly and parameter-free methodology, compression distances, to our specific case by means of a video ASCII format. Particularly, we have used the well-known Normalized Compression Distance (NCD).

In order to use the NCD over video streams we defined a video-to-ASCII conversion methodology. This allows us to make use of compression distances with video objects with successfully results. In this manner, the activity of the video samples is located and casted into text files based on its location in the video frames. Our assumption is that each activity should be expressed with a particular movement signature which, on average, should be shared among various subjects. To corroborate this assumption, we have tested this methodology over different video samples using the *Activities of Daily Living Dataset* [22].

The results presented in this paper show that applying our methodology produces a remarkable clustering along the dataset, which suggests the NCD can be applied to the context of video HAR with success. In the same vein, Fig. 4 shows that the majority of the activities, for this specific database, are fine identified while only a minority are not. This means, that some pairs of activities are too similar to discriminated which videos belong to each activity using this analysis. With this approach, we achieved reasonable results without taking in consideration the particularities of the dataset.

As future work we plan to test and improve our new format over different data sets. In the same vein, we intend to produce alternative video-to-ASCII formats to measure different characteristics of the video activity, and thereby, to add robustness to the system (redundancy). Measuring the vector movement (instead of the activity index) or segmenting the video into multiple ASCII files, are examples of possible alternatives to our method. In summary, we expect to improve the capabilities of the methodology presented in this work exploring different compression algorithms, conversion methodologies and video representations.

Acknowledgment. This work was funded by Spanish project of MINECO/FEDER TIN2014-54580-R and TIN2017-84452-R, (http://www.mineco.gob.es/).

References

1. Akkaladevi, S.C., Heindl, C.: Action recognition for human robot interaction in industrial applications. In: 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), pp. 94–99, November 2015
2. Avgerinakis, K., Briassouli, A., Kompatsiaris, I.: Recognition of activities of daily living for smart home environments. In: 2013 9th International Conference on Intelligent Environments, pp. 173–180, July 2013
3. Beauchemin, S.S., Barron, J.L.: The computation of optical flow. ACM Comput. Surv. **27**(3), 433–466 (1995)
4. Bux, A., Angelov, P., Habib, Z.: Vision based human activity recognition: a review. In: Angelov, P., Gegov, A., Jayne, C., Shen, Q. (eds.) Advances in Computational Intelligence Systems. AISC, vol. 513, pp. 341–371. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-46562-3_23
5. Cerra, D., Datcu, M.: Expanding the algorithmic information theory frame for applications to earth observation. Entropy **15**(1), 407–415 (2013)
6. Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F.: A review on vision techniques applied to Human Behaviour Analysis for Ambient-Assisted Living. Expert. Syst. Appl. **39**(12), 10873–10888 (2012)
7. Chen, T.C., Dick, S., Miller, J.: Detecting visually similar web pages: application to phishing detection. ACM Trans. Internet Technol. **10**(2), 5:1–5:38 (2010)
8. Cilibrasi, R., Vitanyi, P.M.B.: Clustering by compression. IEEE Trans. Inf. Theory **51**(4), 1523–1545 (2005)
9. Cilibrasi, R., Cruz, A.L., de Rooij, S., Keijzer, M.: CompLearn Home. CompLearn Toolkit. http://www.complearn.org/
10. Cohen, A.R.: Extracting meaning from biological imaging data. Mol. Biol. Cell **25**(22), 3470–3473 (2014)
11. Cohen, A., Bjornsson, C., Temple, S., Banker, G., Roysam, B.: Automatic summarization of changes in biological image sequences using algorithmic information theory. IEEE Trans. Pattern Anal. Mach. Intell. **31**(8), 1386–1403 (2009)
12. González-Pardo, A., Granados, A., Camacho, D., de Borja Rodríguez, F.: Influence of music representation on compression-based clustering. In: IEEE World Congress on Evolutionary Computation, pp. 2988–2995 (2010)
13. Granados, A., Cebrian, M., Camacho, D., de Borja Rodriguez, F.: Reducing the loss of information through annealing text distortion. IEEE Trans. Knowl. Data Eng. **23**(7), 1090–1102 (2011)

14. Granados, A., Koroutchev, K., de Borja Rodríguez, F.: Discovering data set nature through algorithmic clustering based on string compression. IEEE Trans. Knowl. Data Eng. **27**(3), 699–711 (2015)
15. Gueguen, L., Datcu, M.: A similarity metric for retrieval of compressed objects: application for mining satellite image time series. IEEE Trans. Knowl. Data Eng. **20**(4), 562–575 (2008)
16. Guha, T., Ward, R.K.: Image similarity using sparse representation and compression distance. IEEE Trans. Multimed. **16**(4), 980–987 (2014)
17. Khan, Z.A., Sohn, W.: Abnormal human activity recognition system based on R-transform and kernel discriminant technique for elderly home care. IEEE Trans. Consum. Electron. **57**(4), 1843–1850 (2011)
18. Lavesson, N., Axelsson, S.: Similarity assessment for removal of noisy end user license agreements. Knowl. Inf. Syst. **32**(1), 167–189 (2012)
19. Li, M., Chen, X., Li, X., Ma, B., Vitanyi, P.: The similarity metric. IEEE Trans. Inf. Theory **50**(12), 3250–3264 (2004)
20. Liu, M., Chen, C., Liu, H.: Time-ordered spatial-temporal interest points for human action classification. In: 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 655–660, July 2017
21. Maddalena, L., Petrosino, A.: A self-organizing approach to background subtraction for visual surveillance applications. IEEE Trans. Image Process. **17**(7), 1168–1177 (2008)
22. Messing, R., Pal, C., Kautz, H.: Activity recognition using the velocity histories of tracked keypoints. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 104–111, September 2009
23. Qiao, L., Nahrstedt, K.: Comparison of MPEG encryption algorithms. Comput. Graph. **22**(4), 437–448 (1998)
24. Roitberg, A., Perzylo, A., Somani, N., Giuliani, M., Rickert, M., Knoll, A.: Human activity recognition in the context of industrial human-robot interaction. In: 2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1–10, December 2014
25. Sarasa, G., Granados, A., Rodriguez, F.B.: An approach of algorithmic clustering based on string compression to identify bird songs species in xeno-canto database. In: 2017 3rd International Conference on Frontiers of Signal Processing (ICFSP), pp. 101–104, September 2017
26. Wang, X.: Intelligent multi-camera video surveillance: a review. Pattern Recognit. Lett. **34**(1), 3–19 (2013)
27. Wu, S., Oreifej, O., Shah, M.: Action recognition in videos acquired by a moving camera using motion decomposition of Lagrangian particle trajectories. In: 2011 International Conference on Computer Vision, pp. 1419–1426, November 2011
28. Yan, Y., Ricci, E., Liu, G., Sebe, N.: Egocentric daily activity recognition via multitask clustering. IEEE Trans. Image Process. **24**(10), 2984–2995 (2015)
29. Yu, T., Wang, Z., Yuan, J.: Compressive quantization for fast object instance search in videos. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 726–735, October 2017
30. Zhang, G., Chanson, H.: Application of local optical flow methods to high-velocity free-surface flows: validation and application to stepped chutes. Exp. Therm. Fluid Sci. **90**, 186–199 (2018)
31. Zhang, S., Wei, Z., Nie, J., Huang, L., Wang, S., Li, Z.: A review on human activity recognition using vision-based method. J. Healthc. Eng. **2017** (2017)