Junsong Yuan and
Zicheng Liu

# TechWare: Video-Based Human Action Detection Resources

Please send suggestions for Web resources of interest to our readers, proposals for columns, as well as general feedback, by e-mail to Dong Yu ("Best of the Web" associate editor) at dongyu@microsoft.com.

In this issue, "Best of the Web" focuses on video-based human action detection, which has recently been demonstrated to be very useful in a wide range of applications including video surveillance, telemonitoring of patients and senior people, medical diagnosis and training, video content analysis and search, and intelligent human computer interaction. As video camera sensors become less expensive, this approach is increasingly attractive since it is low cost and can be adapted to different video scenarios.

Actions can be characterized by spatiotemporal patterns. Similar to the object detection, action detection finds the reoccurrences of such spatiotemporal patterns through pattern matching. Different from action recognition or categorization, where each action video is classified into one of the predefined action classes, the task of action detection needs to identify not only which type of actions occurs but also where (spatial location in the image) and when (temporal location) it occurs in the video. In general, it is a more challenging problem than action categorization. On the other hand, compared with human motion capture, which requires recovering the full pose and motion of the human body, the task of action detection only requires detecting the occurrences of a certain type of actions.

Digital Object Identifier 10.1109/MSP.2010.937496

This column first reviews the general human action detection technology including popular video features and detection methods. It then discusses some useful online resources. Unless otherwise noted, the resources are free.

## VIDEO FEATURES FOR ACTION DETECTION

The development of video-based action detection technology has been ongoing for decades. The extraction of appropriate features is critical to action detection. Ideally, visual features are able to handle the following challenges for robust action detection: i) viewpoint variations of the camera, ii) performing speed variations for different people, iii) different anthropometry of the performers and their movement style variations, and iv) cluttered and moving backgrounds.

In the early days, human bodies were tracked and segmented from the videos to characterize actions and motion trajectories are popularly used to represent and recognize actions. Unfortunately, only limited success has been achieved because robust object tracking is itself a nontrivial task. Recently, interest point-based video features show promising results in the action detection research. Such interest point-based video features do not require foreground/background separation or human tracking.

In the following, we list the resources for four types of interest-point based features. The first is called space-time interest point (STIP), which is developed by Laptev and Lindeberg. STIP features have been frequently used for action recognition. However, the detected interest points are usually quite sparse, and it is time consuming to extract STIP features for high-resolution videos.

The second type of interest point features is named dense and scale-invariant spatiotemporal interest point (DSI-STIP), which is developed by Willems et al. Compared to the STIP features, the DSI-STIP features are scale-invariant (both spatially and temporally) and densely cover the video. The feature extraction is accelerated through the use of approximate box-filter operations on an integral video structure.

The third type of interest point features is called sparse spatiotemporal feature, developed by Dollar et al. The sparse spatiotemporal features are usually denser than the STIP features. However, they do not contain features at multiple scales. In addition, as the goal mainly focuses on local feature detection, the proposed feature descriptor is relatively simple. For example, compared with the histogram of gradient and histogram of flow descriptor proposed by Laptev and Lindeberg, it does not distinguish the appearance and motion features.

The fourth type of interest point features is called 3-D SIFT, developed by Scovanner et al. This descriptor is similar to scale invariant feature transformation (SIFT) descriptor except that the gradient direction for each pixel is a three-dimensional vector. It can work with any interest point detector.

### *STIP DETECTOR*

*http://www.irisa.fr/vista/Equipe/People/Laptev/download.html*

### *DSI-STIP DETECTOR*

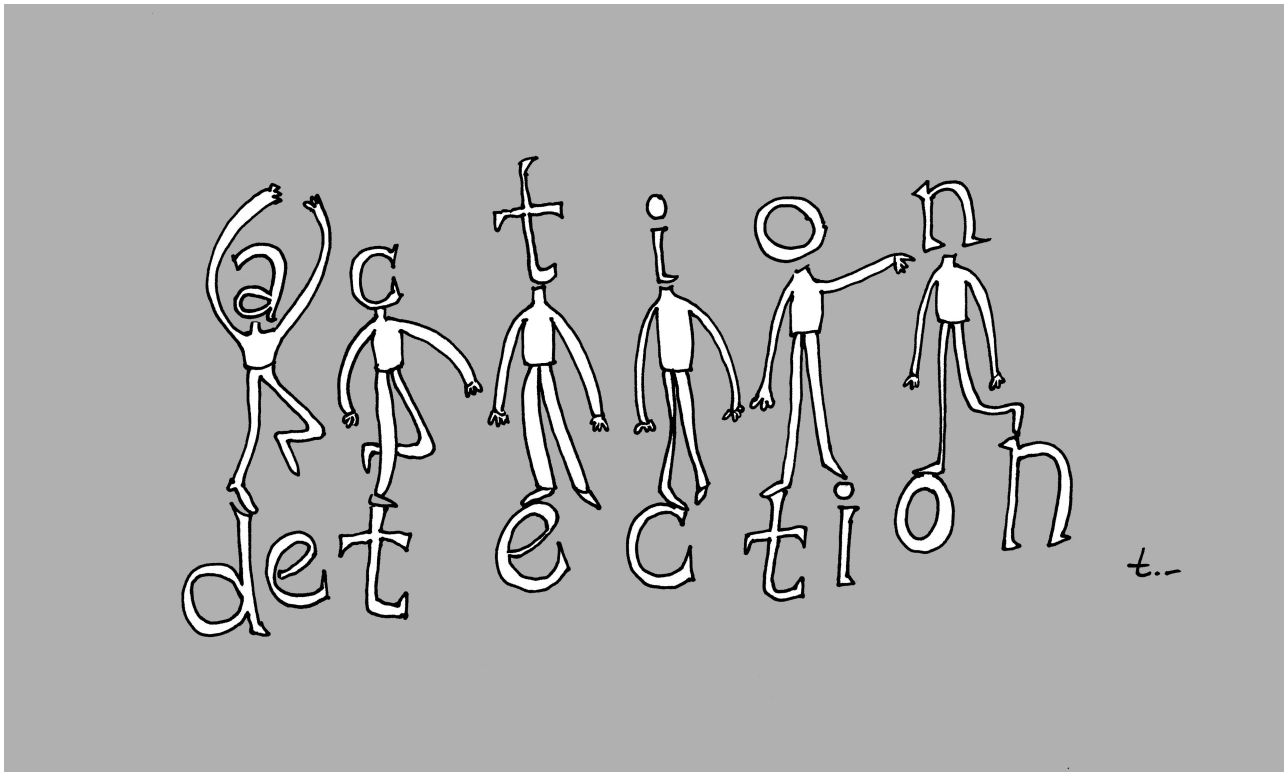
*http://homes.esat.kuleuven.be/~gwillems/research/Hes-STIP/*

### *SPARSE SPATIOTEMPORAL FEATURES*

*http://vision.ucsd.edu/~pdollar/*



1053-5888/10/$26.00©2010IEEE




**Video-based human action detection. Cartoon by Tayfun Akgul (tayfun.akgul@ieee.org).**

***3-D SIFT***

*http://www.cs.ucf.edu/vision/public_html/source.html*

## ACTION DATABASE

There are many public data sets for human action recognition and detection. We categorize them into two classes: 1) data set for action recognition and 2) data set for action detection. For action recognition, each video clip contains only one action and the task is to classify which type of action it exhibits. On the other hand, for the task of action detection, each video sequence contains multiple actions, and it is required to locate every action and tell where and when a specific action happens.

***DATA SET FOR ACTION RECOGNITION:***

*http://www.cs.rochester.edu/~rmessing/uradl*

*[Rochester daily living data set]*

The Rochester daily living data set includes daily actions such as answering phone, chopping banana, dialing phone, drinking water, eating banana, eating snack, lookup in phone, peeling banana, using silverware, and writing on white board. Code source is available to extract and track the STIP features and to generate a histogram of velocity-history code words.

*http://www.nada.kth.se/cvap/actions/*

*[KTH action recognition data set]*

The KTH action recognition data set contains six types of actions: hand waving, hand clapping, boxing, walking, jogging, and running. Each video clip contains one action only. This data set is a popular data set for action categorization benchmark.

*http://www.wisdom.weizmann.ac/il/~vision/SpaceTimeActions.html*

*[Weitzmann data set]*

The Weizmann database contains 90 low-resolution (180 x 144, 50 f/s) video sequences showing nine different people, each performing ten actions including run, walk, skip, jumping jack, jump-forward-on-two-legs, jump-in-place-on-two-legs, gallopsideways, wave-two-hands, waveone-hand, or bend. This data set is another popular data set for action categorization benchmark.

*http://www.cs.ucf.edu/~liujg/YouTube_Action_dataset.html*

*[Youtube action data set]*

The Youtube action data set contains 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. This is a challenging data set due to the cluttered and moving backgrounds.

*http://server.cs.ucf.edu/~vision/data.html*

*[UCF sports action data set]*

The University of Central Florida (UCF) sports action data set includes a set of sports actions collected from broadcast television channels such as the BBC and ESPN. The video sequences are obtained from a wide range of stock footage Web sites including the BBC Motion gallery, and Getty Images. Actions in this data set include: diving, golf swinging, kicking, lifting, horseback riding, running, skating, swinging, and walking.

### DATA SET FOR ACTION DETECTION:

*http://server.cs.ucf.edu/~vision/data.html*

*[UCF aerial action data set]*

The UCF aerial action data set features video sequences that are obtained using an R/C-controlled blimp equipped with a high-definition camera mounted on a gimbal. The data set contains a number of actions captured at different heights and aerial viewpoints. The actions include: walking, running, digging, picking up an object, kicking, opening a car door, closing a car door, opening a car trunk, and closing a car trunk. These actions are recorded at different flying altitudes that range from 400–450 ft, performed by different actors.

*http://research.microsoft.com/~zliu/ActionRecoRsrc*

*[MSR action data set II]*

The Microsoft Research (MSR) action data set II contains 54 video sequences and has in total 203 action instances of three categories: hand clapping, hand waving, and boxing, performed by many subjects. Each sequence contains multiple types of actions. Some sequences contain actions performed by different people. There are both indoor and outdoor scenes. All of the video sequences are captured with cluttered and moving backgrounds. The ground truth labeling of actions and the detection code are provided.

*http://www.cs.cmu.edu/~yke/video/#Dataset*

*[CMU data set]*

The Carnegie Mellon University (CMU) action data set contains actions such as hand waving, jumping jacks, picking up, and pushing elevator button. There are both indoor and outdoor scenes. All of the video sequences are captured with cluttered and moving backgrounds. The ground truth labeling of actions and the code are provided.

*http://www.irisa.fr/vista/Equipe/People/Laptev/download.html*

*[Hollywood data set]*

The Hollywood human action (HOHA) data set contains two video data sets: Hollywood human actions data set and Hollywood-2 human actions and scenes data set. Hollywood human actions data set contains human action from 32 movies. Each video clip is labeled according to one or more of eight action classes: AnswerPhone, GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp, and StandUp. Hollywood-2 data set contains 12 classes of human actions and ten classes of scenes distributed over 3,669 video clips and approximately 20 hours of video in total.

*http://www.cc.gatech.edu/cpl/projects/socialgames/#_Data*

*[Social game data set]*

The social game data set contains many infant social games such as peek-a-boo and so-big. The task includes the automatic retrieval of social games from laboratory and home videos.

*http://visionpc.cs.uiuc.edu/projects/activity*

*[UIUC action data set]*

The UIUC data set contains two parts: data set one that consists of 532 high resolution sequences of 14 activities performed by eight actors, and data set two that consists of three badminton sequences downloaded from Youtube. The sequences are one single and two double matches at the Badminton World Cup 2006. The source code of the motion context descriptor is provided.

*http://www.itl.nist.gov/iad/mig//tests/trecvid/2009*

*[TRECVID event detection data set]*

The TRECVID event detection data set is intended to help develop advanced technology for event detection in video surveillance. The task is to detect a number of predefined human actions, including PersonRuns, CellToEar, ObjectPut, PeopleMeet, PeopleSplitUp, Embrace, Pointing, ElevatorNoEntry, OpposingFlow, and TakePicture. The video sequences are recorded by airport surveillance cameras.

*http://www.kitware.com/viratdata/*

*[VIRAT data set]*

The VIRAT data set is designed for evaluating the performance of detection and classification of human/vehicle actions, events, and activities in real-world and natural videos. The data set is challenging in terms of resolution and background clutter. It involves diverse data collection.

*http://cvrc.ece.utexas.edu/SDHA2010/index.html*

*[Semantic description of human activities data set]*

This data set was used for the 2010 International Conference on Pattern Recognition Contest on Semantic Description of Human Activities (SDHA 2010). It is composed of three different types of activity recognition tasks: high-level human interaction recognition, aerial view activity classification, and wide-area activity search and recognition. It is a benchmark to test techniques on realistic surveillance-type videos.

## ACTION DETECTION ALGORITHMS

As mentioned earlier, action detection requires an accurate localization of the action in the scene, despite cluttered and dynamic backgrounds and other types of action variations. To address the challenges, different action detection algorithms have been developed recently. We list a few action detection algorithms below, where the codes are accessible.

### SPATIOTEMPORAL SUBVOLUME SEARCH FOR ACTION DETECTION

*http://research.microsoft.com/~zliu/ActionRecoRsrc*

Similar to the sliding window-based object detection, this approach detects actions by searching the reoccurrences of spatiotemporal patterns through pattern matching. It can handle cluttered and dynamic backgrounds as well as action variations. To speed up the search of spatiotemporal subvolumes, a spatiotemporal branch-and-bound search is proposed. The idea of using branch-and-bound search is first proposed for object detection. It is later extended to search actions in video volumes.

The branch-and-bound search avoids the exhaustive checking of all possible bounding boxes/subvolumes while still guarantees that the optimal solution can be found. Although the

worst case complexity is the same as that of the exhaustive search, practically it is much faster than the exhaustive search. The binary code of the branch-and-bound search is downloadable from the Web page.

### VOLUMETRIC FEATURES FOR EVENT DETECTION

*http://www.cs.cmu.edu/~yke/video/*

This method explores the use of volumetric features for event detection. It correlates spatiotemporal shapes to video clips that have been automatically segmented. As it works on over-segmented videos, background subtraction for reliable object segmentation is not required. A flow-based correlation technique is applied for matching, and can detect a wide range of actions in video. It can well handle the cluttered background. However, the detection speed is slow due to the large search space. This approach has relatively limited ability to handle action variations because only one action template is used. The code for feature extraction can be downloaded from the Web site.

### SPACE-TIME SHAPE MATCHING FOR ACTION DETECTION

*http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html*

Human action in video sequences can be seen as silhouettes of a moving torso and protruding limbs undergoing articulated motion. This method regards human actions as three-dimensional shapes induced by the silhouettes in the space-time volume. It is a generalization of a two-dimensional shape-analysis technique to the case of volumetric space-time shapes. The technique exploits Poisson equation solutions to extract space-time features such as local space-time saliency, action dynamics, shape structure, and orientation.

### SUCCESSIVE CONVEX MATCHING FOR ACTION DETECTION

*http://cs.bc.edu/~hjiang/scm/demo.html*

In this approach, human actions are represented as sequences of postures. Each posture is represented as a transformed edge map. Specific actions are detected in a video by matching the time-coupled posture sequences to video frames. The template sequence to video registration is formulated as an optimal matching problem. A successive convex matching scheme is used to improve the matching speed. The demo code for successive convex matching can be accessed from the Web site.

## AUTHORS

*Junsong Yuan* (jsyuan@ntu.edu.sg) is an assistant professor with the School of Electrical and Electronic Engineering at Nanyang Technological University, Singapore.

*Zicheng Liu* (zliu@microsoft.com) is a researcher with Microsoft Research, Redmond, Washington.

[SP]

---

[exploratory DSP] continued from page 135

techniques are working well in machine vision applications, we expect to continue to find good inspiration for what might work well for auditory-image-based machine hearing applications. When we find ideas worth trying, it may be easy to obtain implementations that can be adapted to use the output of our auditory analysis stages. Such repurposing of machine vision systems may provide good leverage in machine hearing research.

## CONCLUSION

The machine hearing field is starting to find its feet. Applications are abundant and many are easy to address with known auditory front ends, combined with known feature extraction and machine learning techniques such as those that have proven successful in analogous applications in machine vision.

The signal processing technology involved is diverse but not too complex. Nonlinear filters, correlators, vector quantizers, and online learning algorithms, are involved in ways that can be initially fairly simple, yet leave room for open-ended research and improvement. Cooperation with researchers in auditory psychology and physiology will be highly valued on both ends.

Curing our machines' deafness, leveraging our knowledge of the amazing capabilities of the mammalian cochlea and auditory brain is a goal that will keep this field busy for a while and that will provide rewards on many fronts.

## AUTHOR

*Richard F. Lyon* (dicklyon@ieee.org) is a research scientist at Google, Inc., and a Fellow of the IEEE.

## REFERENCES

[1] J. Treichler, "Signal processing: A view of the future, Part I," *IEEE Signal Processing Mag.,* vol. 26, no. 2, pp. 116–120, 2009.

[2] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, pp. 128–133, 1951. Reprinted in *Physiological Acoustics,* E. D. Schubert, Ed. Stroudsburg, PA: Dowden, Hutchinson and Ross, Inc., 1979.

[3] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psychol.,* vol. 41, no. 1, pp. 35–39, 1948.

[4] H. Bourlard, H. Hermansky, and N. Morgan, "Towards increasing speech recognition error rates," *Speech Commun.,* vol. 18, no. 3, pp. 205–231, 1996.

[5] L. Watts, "Commercializing auditory neuroscience," in *Proc. Frontiers of Engineering: Reports on Leading-Edge Engineering 2006 Symp.*, 2007, p. 5.

[6] R. F. Lyon, A. G. Katsiamis, and E. M. Drakakis, "History and future of auditory filter models," in *Proc. IEEE Int. Conf. Circuits and Systems*, 2010, pp. 3809–3812.

[7] R. F. Lyon, "Filter cascades as analogs of the cochlea," in *Neuromorphic Systems Engineering: Neural Networks in Silicon*, T. S. Lande, Ed. Norwell, MA: Kluwer, 1998, pp. 3–18.

[8] S. Mandal, S. M. Zhak, and R. Sarpeshkar, "A bio-inspired active radio-frequency silicon cochlea," *IEEE J. Solid-State Circuits*, vol. 44, no. 6, 2009, pp. 1814–1828.

[9] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Proc. 9th Int. Symp. Hearing, Auditory Physiology and Perception*, Y. Cazals, L. Demany, and K. Horner, Eds. Oxford: Pergamon, 1992, pp. 429–446.

[10] M. Slaney and R. F. Lyon, "On the importance of time—A temporal representation of time," in *Visual Representations of Speech Signals,* M. Cooke, S. Beet, and M. Crawford, Eds. New York: Wiley, 1993, pp. 95–116.

[11] D. Grangier and S. Bengio, "A neural network to retrieve images from text queries," in *Proc. Artificial Neural Networks—ICANN 2006,* 2006, pp. 24–34.

[12] M. Rehn, R. F. Lyon, S. Bengio, T. C. Walters, and G. Chechik, "Sound ranking using auditory sparse-code representations," in *ICML Workshop Sparse Methods for Music Audio,* 2009.

[SP]