THETIS: THree Dimensional Tennis Shots A human action dataset

Sofia Gourgari Georgios Goudelis Konstantinos Karpouzis
Stefanos Kollias
National Technical University of Athens
Image Video and Multimedia Systems Laboratory
*

Abstract

The detection and classification of human movements, as a joint field of Computer Vision and Pattern Recognition, is used with an increasing rate in applications designed to describe human activity. Such applications require efficient methods and tools for the automatic analysis and classification of motion capture data, which constitute an active field of research. To facilitate the development and the benchmarking of methods for action recognition, several video collections have previously been proposed. In this paper, we present a new video database that can be used for an objective comparison and evaluation of different motion analysis and classification methods. The database contains video clips that capture the 3D motion of individuals. To be more specific, the set consists of 8374 video clips, which contain 12 different types of tennis actions performed by 55 individuals, captured by Kinect. Kinect provides the depth map of motion data and helps to extract the 3D skeletal joint connections. Performing experiments using state of the art algorithms, the database shows to be very challenging. It contains very similar to each other actions, offering the opportunity to algorithms dedicated to gaming and athletics, to be developed and tested. The database is freely available for research purposes.

1. Introduction

Recognition of human action comprises an important part in the field of computer vision. The aim of this research effort is the automated analysis and recognition of human behavior captured in image sequences (videos).

*This work was co-financed by ICT-Project Siren, under contract (FP7-ICT-258453), by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES. Investing in knowledge society through the European Social Fund.

Video sequence elaboration has progressed from the action detection level, to recognition of actions and interactions as individual facts. Human action recognition by a computer, involves the understanding of human behavior which is a highly complex task. Shape and form of human body cannot be strictly specified due to multiple joints, while clothing can dramatically affect the processing. Lighting variations, external noise (e.g. shadows), view angle, are only few of the major problems.

Understanding of human action can be approached by different levels of examined details, according to the complexity of each case. Modelling and recognition of human behavior presupposes the labelling and classification of the different kind of actions [2]. According to the action complexity, four basic categories of human action can be identified. The first category is the movement of some part of the body, like the rising of a hand (*gestures*). The second category is characterized by the aggregate movements of one person (*actions*). Such actions can be walking, jogging, boxing, etc. *Interaction* is the complex sequence of movements performed by several individuals interfering to each other and may or may not include an object. This comprises the third category while the last one is *group activity* which actually is a set of activities performed by groups of persons. "Queue" is a characteristic example of a group activity.

The applications related to human action recognition are many. Biometrics based on human behavior, surveillance systems, interactive environments and applications, are only a few of them. Since research interest for the specific field is broad, there is a large number of proposed techniques trying to provide solutions to different problems concerning automated activity recognition. In this effort, researchers make use of datasets specially created to serve the exact purpose. Since the problem has many views and concerns a variety of possible applications, different datasets apply to different scenarios. In following, a brief review of some of

the available datasets is provided.

Human action datasets could be sorted into three categories. The first category contains sets that have been designed for the evaluation of general purpose action recognition systems. Probably the most known are the KTH [20] and the Weizmann [3]. The first contains six types of typical human actions (e.g. walking) performed several times by 25 subjects in four different scenarios. The second is a collection of 90 low-resolution video sequences showing nine different people, each performing 10 natural actions (e.g. running).

In [17] collection of subjects performing several actions from different views using a network of 8 embedded cameras is presented. The database contains two sets: the first is proposed for the evaluation of unit actions, while the second for the evaluation of interleaved sequences of actions. In [16] virtual human action silhouette (ViHASi) data is introduced. The specific database is mostly oriented to silhouette-based techniques. In the same category, authors in [21] have collected a large body of human action video (MuHAVi) data using 8 cameras. The set contains 17 actions performed by 17 actors. INRIA Xmas Motion Acquisition Sequences (IXMAS) presented in [26], is a multiview dataset for view-invariant human action recognition applications. The set comprised of 13 daily-live motions performed each 3 times by 11 actors. The actors choose freely position and orientation.

The datasets of the second category are mainly oriented to application that arise from realistic environments (e.g. airports). An example of this category is presented as a challenge in [5]. The challenge aims to motivate researchers to explore techniques that may address the issues with recognizing human actions in low-resolution videos. Usually such videos are the ones filmed from a distance view such as aerial videos. The videos show a single person, performing various actions taken from the top of a building and the average height of human figures is about 20 pixels. The idea behind this set is to be a test object for military applications and surveillance systems where distance views is a common case.

Moving in a similar concept the same lab in [19] provide another set to motivate researchers to explore the recognition of complex human activities from continuous videos, taken in realistic settings. Each video contains several human-human interactions (e.g. hand shaking) occurring sequentially and/or concurrently. Authors in [7], present a collection of realistic scenarios in a multi-camera network environment (VideoWeb) involving multiple persons performing dozens of different repetitive and non-repetitive activities.

In another approach [24] aiming to applications developed for every day common activities, a set called "TUM Kitchen Data Set" is presented as a comprehensive collection of activity sequences recorded in a kitchen environment equipped with multiple complementary sensors. The recorded data consists of observations of naturally performed manipulation tasks as encountered in everyday activities of human life.

In [6] a multiview human action dataset is provided. The set with name UCF-ARG (University of Central Florida-Aerial camera, Rooftop camera and Ground camera) Data, consists of 10 actions performed by 12 actors recorded from a ground camera, a rooftop camera at a height of 100 feet and an aerial camera mounted onto the payload platform of an aerostat helium balloon.

The LIRIS human activities dataset [27] contains (gray/rgb/depth) videos showing people performing various activities taken from daily life (discussing, telephone calls, giving an item etc.). The samples are given in two subsets recorded by Microsoft's Kinect and a Sony consumer camcorder respectively. G3D also recorded using Kinect in [4], contains a range of gaming actions providing synchronized data (video, depth and skeleton). The set contains 10 subjects performing 20 gaming actions recorded in a controlled indoor environment with a fixed camera as a typical gesture based gaming setup.

Finally, the third category comprises of a number datasets that have been accrued from the collection of existent videos, received from media, television programs and movies. Such an example is the dataset presented in [12] which intends to provide a comprehensive benchmark for human action recognition in realistic and challenging settings. The dataset is composed of video clips from 69 movies.

In [14] an interaction dataset is presented that consists of 300 video clips collected from over 20 different TV shows and containing 4 interactions: hand shakes, high fives, hugs and kisses, as well as clips that don't contain any of the interactions. In [13], TV Human Interactions dataset consists of 300 video clips collected from over 20 different TV shows and containing 4 interactions: hand shakes, high fives, hugs and kisses, as well as clips that don't contain any of the interactions.

In [11] a set of 11 action categories for recognition of realistic actions from videos "in the wild" is presented. The challenge that the specific set provides, is how to extract reliable and informative features from the unconstrained videos. The set is mainly collected from Youtube and contains video samples with large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc.

The same institute, presents in [18] another dataset that consists of a set of actions collected from various sports which are typically featured on broadcast television channels such as the BBC and ESPN. The video sequences were obtained from a wide range of stock footage websites in-

cluding BBC Motion gallery, and GettyImages. A third approach for another set by the same institute, is given in [23] named UCF101 and is an action recognition data set of realistic action videos, collected from YouTube, having 101 action categories. This data set is an extension of a previous published set named UCF50 that contained 50 action categories.

Another approach is presented in [8], where a video database of actions and a full testing protocol for studying action similarity from videos are provided. The idea is to stop learning the properties of particular action classes and start to gain a more principled understanding of what makes actions different or similar. The named ASLAN set (Action Similarity Labeling Chalenge) includes thousands of videos collected from the web, in over 400 complex action classes. The benchmark protocol is mainly focused on action similarity than action classification and testing is focused on *never-before-seen* actions.

In this paper, we present a sport based human action dataset comprised of the 12 basic tennis shots. Although the specific shots may be distinguishable to an expert's eye, some of them can be quite confusing to someone irrelevant to tennis. This provides an extra challenge to automated action recognition systems. The shots have been performed by experts and amateurs rising this way the difficulty of the task. The data are provided in 5 different synchronized forms (RGB, silhouettes, depth, 2D skeleton and 3D skeleton)[1].

2. THree dimEnsional Tennis Shots (THETIS)

In this section we describe the camera setup, the database content and the processing steps followed in order to derive THETIS database. The set is comprised of a set of 12 basic Tennis shots performed by 31 amateurs and 24 experienced players. Each shot has been performed several times resulting in 8734 videos of the AVI format. The total duration of the videos is 7 hours and 15 minutes. The shots that have been performed are given bellow:

- Backhand with two hands
- Backhand
- Backhand slice
- Backhand volley
- Forehand flat
- Forehand open stands
- Forehand slice
- Forehand volley
- Service flat
- Service kick
- Service slice
- Smash

[1]The dataset can be accessed via: http://thetis.image.ece.ntua.gr/

2.1. Camera set up and capture conditions

As already mentioned, the recording device for the proposed dataset was Microsoft's Kinect which was initially created as a webcam-style add-on peripheral for the XBOX 360 game console. It enables users to control and interact with the game console without the need to touch a game controller. Kinect uses an infrared projector, a camera and a special microchip to track the movement of objects and individuals in three dimensions.

For the communication of the specific device with the computer, the open source software OpenNI [1] and the middleware of PrimeSence [15] that contains the driver for the 3D sensor were used. The OpenNI framework is an open source SDK used for the development of 3D sensing middleware libraries and applications. To be more specific, it achieves the communication with the optoacoustic sensors and analyzes the optic and acoustic data that the device records.

The recordings of the proposed dataset were performed into two different indoor places. The reason for making indoor recordings was Kinect's limitation to perform under direct sunlight. More specifically, since Kinect projects and reads an infrared mesh to construct 3D information, is highly affected by the infrared radiation of the corresponding sunlight spectrum.

The capture device was set at 1.6 meters high from the ground and remained static during every capture. The action performing point was set to around 1.5 meters, while every action was repeated several times from the same ground point. At this point we should mention that before they perform a shot, amateur players attend a demonstration of every action by a tennis instructor. In following, they try to follow suit the instructions.

As far as it concerns the background, it does not remain static since most of the times varies from one shot to the other, containing multiple persons acting in different ways behind the action scene (passing, playing basketball etc.). There may also be differences in view angle but are judged as unimportant. Examples of background variations are illustrated in Figure 1.

2.2. Structure

Initially, each person repeats each of the above 12 tennis shots 3 to 4 times. This resulted in 660 files of ONI type. Since retention of the relevant data in ONI format

Figure 1. Samples of background variations in THETIS.

files would be qualificatory for wide use, as OpenNI application would be required, conversion of the ONI files to a widely spread format was necessary. Thus, all files have been conversed to AVI format using an algorithm based on OpenNI, developed in our lab. The specific application offers the following features: Isolation of depth data recorded by Kinect's depth sensor. Extraction of the silhouette of the person performing the action. Extraction of the skeleton of the human body through detection of body joints. Illustration of the skeleton relevant data to 2 and 3 dimensions. From every ONI file, 5 synchronized AVI files have been produced:

- An AVI file that contains the RGB information
- An AVI file that contains the depth information
- An AVI file that contains the silhouette of the person
- An AVI file that contains the movement of the skeleton in 2 dimensions
- An AVI file that contains the movement of the skeleton in 3 dimensions

This results in 3300 AVI files that in following are manually cropped in single actions. The aim of the above described procedure is to receive from each initial video, 3 new ones containing a single repetition of each shot. Thus, 1980 RGB videos, 1980 depth videos and 1980 mask videos (silhouette) have been produced.

As far as 2D and 3D skeletons are concerned, these are not always provided for all of the repetitions. This is due to the limitations appear when one is trying to obtain the skeletons from the initial ONI file. More specifically, each performer has to initially take a calibration pose at the start of the recording. Failing to do so, results in skeleton obtainment failure. Unfortunately, calibration pose was not successful in a few cases and this is something that could have

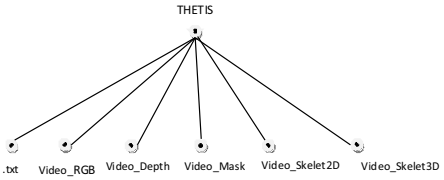

Figure 2. THETIS' structure.

not been checked in advance. Additionally, some of the participants performed some of the shots in extreme high speed extracting in this way the skeleton, only to later repetitions where the system has managed to calibrate. The number of skeleton videos finally extracted, was 1217 for the 2 dimensions and 1217 for the 3 dimensions respectively. The folder structure of the dataset is given in Figure 2.

The summary of the contents of each folder is given below:

- Video RGB: contains 1980 AVI files
- Video Depth: contains 1980 AVI files
- Video Mask: contains 1980 AVI files
- Video Skelet2D: 1217 AVI files
- Video Skelet3D 1217 AVI files
- txt: detailed description of the contents of each folder of the dataset

Snapshots of all video types of a participant performing a tennis shot (*forehand slice*), are illustrated in Figure 3.

3. Experimental Results

To evaluate the potentials and the challenge that THETIS offers for future research, we performed a set of experiments applying two state of the art algorithms for human action recognition. For the specific purpose, we used all skeleton based videos aligned to 3 dimensions and the videos representing the depth information as these are the 3D relevant ones.

Additionally, we performed the same type of experiments using the same protocol on a common known and widely used dataset, the so-called KTH. The purpose of the specific experimental procedure and the comparison with a well-known database was not to relegate the value of the algorithms used, but to highlight the challenges arising by the proposed database.

A testified effective way for detection of important points within an image and subsequently within a video, is to find *points of interest*. The last decade, a number of techniques

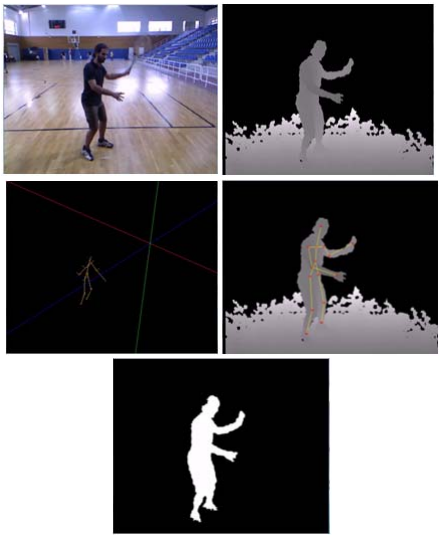

Figure 3. Snapshots of all video types of a participant performing the *forehand slice* shot.

have been proposed trying to find such spacio-temporal points that are consistent in scaling changes of human actions [2].

In the described experimental procedure, for the detection of the spatio-temporal points of interest and the extraction of their descriptors we followed the method described in [10] and we used the the code provided by the authors. The code is an extension of Harris 3D detector proposed in [9] which detects spacio-temporal points of interest and calculates the spacio-temporal descriptors, Histograms of Oriented Gradient (HOG) and Histograms of Optical Flow (HOF). Following the same procedure as in [10], we use the same version of the code.

In the second stage, we performed experiments using another human action recognition algorithm, proposed in [25], the so-called Dense Trajectories. The specific method is based in the dense sampling of points detected in each frame and tracks their shift according to the information received by optical flow fields. The number of points can be easily multiplied if optical flow fields have been calculated without cost. Thus, the dense trajectories of the points, finally describe the movement within a video. An example of Dense Trajectories extracted from a video of THETIS database is illustrated in Figure 4.

Since the above feature extraction techniques produce large-size data, a bag of features technique is applied. In other words a visual vocabulary is created to quantize the descriptors. The visual vocabulary is created using K-means [22]. For every visual vocabulary created for each of the descriptors using k-means, a k=500 was used.

For the classification, 12 non-linear SVMs (as many as the classes) have been used. While the leave-one-person-out cross validation approach was used to evaluate performance. The specific protocol was chosen due to its popularity among researches. The specific technique preserves one person's samples (videos) as training set, while the rest of them are used for training. The procedure is repeated N times where N is the number of subjects within the dataset. Performance is reported as the average accuracy of N iterations. Thus, for the KTH set which contains 24 persons, the procedure was repeated 24 times while for the 3D skeleton and depth sets from the THETIS database, the whole procedure was repeated 55 times. The results produced for the different kind of data and descriptors are presented in Tables 1 and 2.

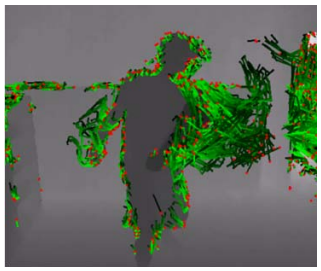

Figure 4. Dense Trajectories extracted from a depth video from THETIS dataset.

Dataset	Average Accuracy (%)
THETIS Depth	60.23
THETIS Skelet3D	54.40
KTH	92.99

Table 1. STIP method for the combination of HOG and HOF descriptors for the 3 different datasets.

Dataset	Average Accuracy (%)		
Descriptor	Trajectory	MBH	combination
TH. Depth	51.59	54.32	57.50
TH. Skelet3D	46.84	50.78	53.08
KTH	86.98	92.32	90.65

Table 2. Dense Trajectory method results using different descriptor combination: Trajectory, MBH, and combination of Trajectory, HOG, HOF and MBH respectively.

4. Conclusion

In this paper we presented THETIS, a new database containing a large set of tennis shots. THETIS combines a number of advantages that may benefit future works on the field of human action recognition. The recordings have been made using Kinect providing this way not only 2D information (RGB and silhouettes), but also depth and extracted 2D and 3D skeletons. The shots have been performed by 31 amateurs and 24 experienced players resulting in a total

of 8734 videos of the 5 types. Our intention is THETIS to set a useful tool not only for tennis related applications, but also for the development of different 2D or 3D based action techniques, sport applications, gaming and others. Finally, it could be a tool for an alternative approach where a system not only recognizes the shot, but also evaluates the expertise of the players.

References

[1] Center for Research in Computer Vision. UCF-ARG Data Set. http://crcv.ucf.edu/data/UCF-ARG.php, 2013. [Online; accessed 17-April-2013].

[2] OpenNI. The standard framework for 3D sensing. http://www.openni.org, 2013. [Online; accessed 17-April-2013].

[3] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, Apr. 2011.

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pages 1395–1402, 2005.

[5] V. Bloom, D. Makris, and V. Argyriou. G3d: A gaming action dataset and real time action recognition evaluation framework. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 7–12, 2012.

[6] C.-C. Chen, M. S. Ryoo, and J. K. Aggarwal. UT-Tower Dataset: Aerial View Activity Classification Challenge. http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html, 2010.

[7] G. Denina, B. Bhanu, H. Nguyen, C. Ding, A. Kamal, C. Ravishankar, A. Roy-Chowdhury, A. Ivers, and B. Varda. Videoweb dataset for multi-camera activities and non-verbal communication. In *Distributed Video Sensor Networks*, pages 335–347. Springer London, 2011.

[8] O. Kliper-Gross, T. Hassner, and L. Wolf. The action similarity labeling challenge. *IEEE Transactions of Pattern Analysis and Machine Intelligence (TPAMI*, 2012.

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

[10] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition (CVPR)*. IEEE, 2008.

[11] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild, 2009.

[12] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[13] A. Patron, M. Marszalek, A. Zisserman, and I. Reid. High five: Recognising human interactions in tv shows. In *Proceedings of the British Machine Vision Conference*, pages 50.1–50.11. BMVA Press, 2010.

[14] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman. Structured learning of human interactions in tv shows. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(12):2441–2453, 2012.

[15] PrimeSense. What is all about. http://www.primesense.com/, 2013. [Online; accessed 17-April-2013].

[16] H. Ragheb, S. A. Velastin, P. Remagnino, and T. Ellis. Vihasi: virtual human action silhouette data for the performance evaluation of silhouette-based action recognition methods. In *VNBA*, pages 77–84, 2008.

[17] S. Ramagiri, R. Kavi, and V. Kulathumani. Real-time multi-view human action recognition using a wireless camera network. In *Distributed Smart Cameras (ICDSC), 2011 Fifth ACM/IEEE International Conference on*, pages 1–6, 2011.

[18] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[19] M. S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). http://cvrc.ece.utexas.edu/SDHA2010/Human Interaction.html, 2010.

[20] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *In Proc. ICPR*, pages 32–36, 2004.

[21] S. Singh, S. A. Velastin, and H. Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *AVSS*, pages 48–55, 2010.

[22] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, Oct. 2003.

[23] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[24] M. Tenorth, J. Bandouch, and M. Beetz. The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1089–1096, 2009.

[25] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3169–3176, 2011.

[26] D. Weinland and E. Boyer. Action recognition using exemplar-based embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage*, pages 1–7, 2008.

[27] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E. Dellandre'a, C.-E. Bichot, C. Garcia, and B. Sankur. The LIRIS Human activities dataset and the ICPR 2012 human activities recognition and localization competition. Technical Report RR-LIRIS-2012-004, LIRIS UMR 5205 CNRS/INSA de Lyon/Universite' Claude Bernard Lyon 1/Universite' Lumie're Lyon 2/E'cole Centrale de Lyon, Mar. 2012.