Review on Vision based Human Activity Analysis

Sreeja Sankaran Nampoothiri
PG Scholar
Dept. Of Electronics and Communication
Vimal Jyothi Engineering College, Chemperi

Anoop B.K
Assistant Professor
Dept. Of Electronics and Communication
Vimal Jyothi Engineering College, Chemperi

ABSTRACT
Recognizing human actions are important in various real time applications. Review on human activity analysis is provided in three sections. The first section in this paper presents an overall classification to Human activity analysis from feature extraction to recognition systems. In the second section a survey is included which provides technical information to activity analysis. Finally a brief description of databases which came across in survey is also included. The overall purpose of this paper is to provide a basic understanding to human activity analysis and to analyze the major challenge in human activity analysis.

General Terms
Human Activity Recognition (HAR), Feature Extraction, Feature Representation, Classification, Recognition

Keywords
Background Subtraction (BS), Human Activity Recognition (HAR), Motion Energy Image (MEI), Hidden Markov Model (HMM), State Vector Machine (SVM)

1. INTRODUCTION
Human activity analysis and recognition is vast and challenging research topic in the field of computer vision and pattern recognition. It has wide variety of applications in real life applications. Applications in surveillance systems, video browsing and human-computer interfaces, robotics, smart phone based mobile applications, content based video search, human-machine interaction , video based intelligent systems, augmented reality , healthcare are the major areas where human activity is of major concern. Even though successful researches have been done to recognize simple human activities, there are lot of constraints in real time environment recognition analysis and higher level of application needs. This survey organizes the human activity recognition and understands the major focus and challenges of research area under human activity analysis.

There are many methods in which human activity can be recognized (1) By Sensors which measures acceleration, vibration, rotation etc. (2) By analyzing radar signals, and (3) Vision based human activity identification from Videos, still images and thermal infrared images used by Bhanu et. al [5] etc (4) Wi-Fi signals. Here we deal with only vision based activity recognition system. Human attention in vision based system is of least importance thus adding an advantage to the same. Activity analysis addresses solutions for activity detection and tracking of humans to person identification. Activity analysis takes at different levels like (1) Part based analysis (2) Single person activities (3) Interaction between two individuals (4) Group Activities and Interactions.

This paper is mainly divided into three sections. In section 2 we shall see the generalized stages for human action recognition which will provide an insight to different stages and its classifications. The next section presents a survey on the latest proposals that has been used for activity analysis. The third section describes on available datasets in use.

2. OVERVIEW OF HUMAN ACTIVITY ANALYSIS
A general overview of human activity analysis is shown in Fig.1.There are three basic classical steps to human activity analysis or recognition.

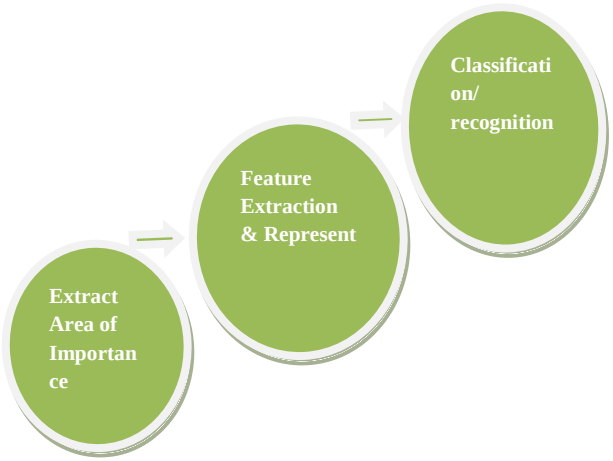

Fig 1: Overview of human activity analysis

- Extraction the area of importance from provided input source
- Feature Extraction and representation
- Classification or recognition

Extracting the area of importance is one of the crucial tasks as the final result of recognition depends on what we extract from a sequence of images or videos and so the one of the challenging task lies in this stage. The area of importance includes human area or portion of particular interest. Features such as height, velocity, motion volumes etc. is then extracted from the area of interest and mostly represented in vector form. Based on the feature vector a trainer will be trained and finally a classifier or different algorithms are used to recognize the activity in the input source. In the coming sections we shall see a detailed description of the three basic steps.

2.1 Extracting Area of Importance/target

This is the initialization procedure for analysis. There are different sources available from which we can change input into desired forms for extracting features. For example MEI and MHI are derived image templates for motion feature extraction. Input sources include infrared images, YouTube videos, surveillance system data, monocular images [3] etc. Extraction of target can be done by

Extraction of human area: Extraction of human area includes silhouettes and human blobs development. Background subtraction method is used to develop shape descriptors such as silhouettes and human blobs. This method is used by Wang et.al [4] and Jenn et. al.[6]. In order for back ground subtraction, background modeling should be done by methods such as kernel density estimation technique [4] or by Gaussian mixture models.

Extraction of Salient region: In order to extract salient/important features some methods like commonly used are STIP method [8], sparse coding [7] and STFT. Here no extraction of human area done. This method only selects the points of interest and then provides Spatio-temporal descriptors. STIP based methods are commonly seen popular and work with low resolution inputs.

2.2 Feature Extraction and Representation

This second stage decides what features to be extracted based on the purpose of application and its representation. From the survey it is seen that this is an important area being focused in action analysis currently. Features can be classified as local features and Global features. Local Features are those referencing to a single patch. Classical interest point based methods help in extracting local features [12]. But global features are those which are considered based on sequence of extracted areas. MHI [19, 4] provide global features.

Type of features extracted includes (1) shape (2) kinematic and (3) Spatio-temporal features. From the shape, shape feature parameters such as height, width etc can be calculated. Kinematics feature parameters include velocity of trajectory, acceleration etc. In case of Spatio-temporal feature include both space time features like spatiotemporal curvature to calculate discontinuity in velocity [11].

After extracting the desired feature parameters, it is necessary to represent each activity uniquely for further processing. The feature parameter initially will of high dimensional representation. Considering the computational speed, memory constraints etc it is necessary to narrow down the high dimensional representation of feature parameters to low dimension. SAX [11], PCA-STOP [20], Linear discriminative analysis, Non negative matrix factorization, Bag-of-Words [7-8] are some examples for reduction to low dimension. The final representation of each action will be a feature vector.

Feature extraction and representation can be shown as in Fig. 2

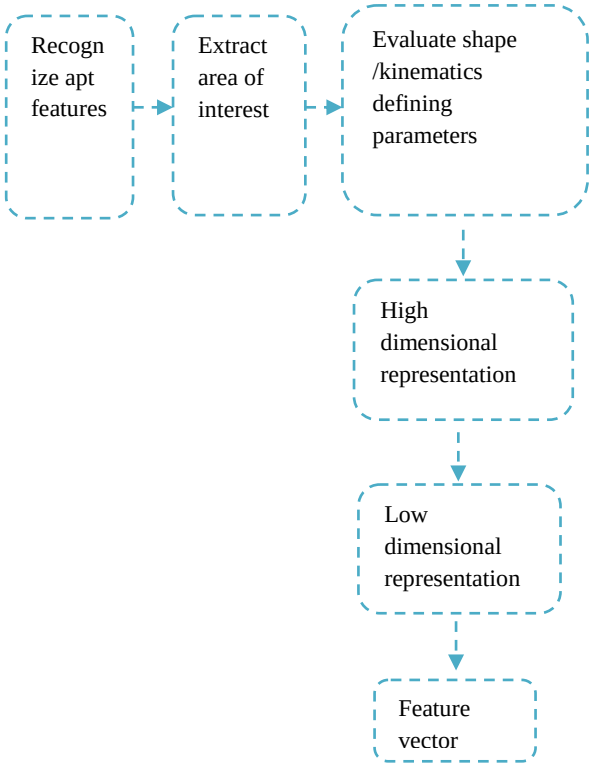

Fig 2: Stages in feature extraction and representation

2.3 Recognition of Human Activity

The final stage after proper activity representation is recognition. Not many developments have been done recently on recognition methods. After feature representation and before recognition there is an intermediate training stage. The set of trained models mostly in form of feature vector is stored in a code book or database. This trained model is used in order to compare the feature vector of unknown activity source with that of trained model eg. comparison by Euclidean distance method. Human action recognition can be classified from different perspectives. Here we will see some of those perspectives for action recognition.

Human action interpretation can be classified into two main categories based on initial extraction. (1) Model Based Approach and (2) Model Less approach [3].

Model Based Approach: Model based approach attempts to recover structural information of human body by representing them as stick or kinematic tree model. Model based approach is usually used in pose extraction. It can use either 3d shape model or 2d shape models. One advantage is that activity parameters can be directly derived from the model .In many situations model based approach is complex in nature.

Model Less approach: does not require construction of any models. They transform high dimensional features to lower dimension representation. These methods usually require training models to recognize the features. They take into consideration the shape of silhouette or motion information of body parts. The disadvantage of model less approach is that it need diverse training samples and needs correct transformation mapping.

The second classification is based upon action interpretation. Hence HAR can be classified as (1) shape based HAR [4] (2)

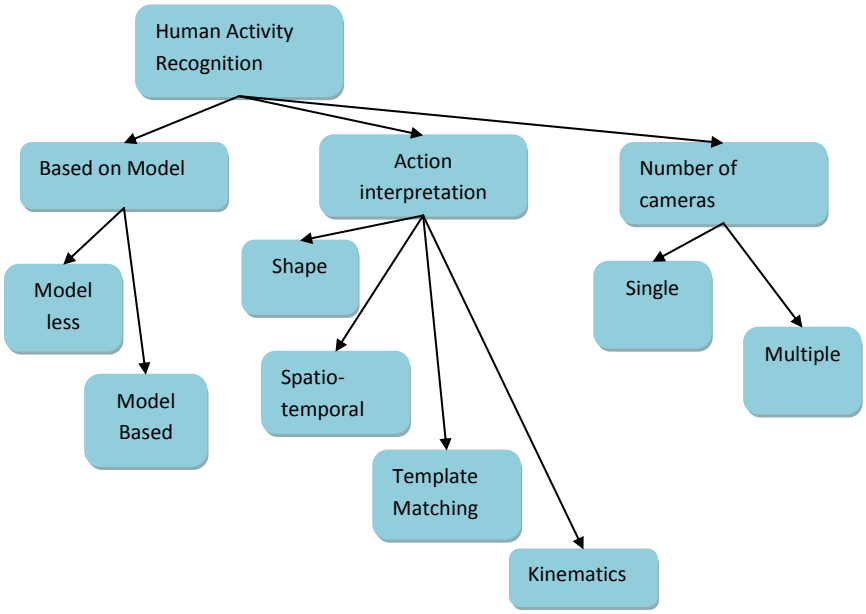

Fig 3: Different classifications for Human activity recognition

Kinematics /Motion Information based HAR (3) Spatio-temporal volumes [6-8] (4) Template Matching. Shape information provides local features such as height , width etc. Kinematics information can be derived from Shape, Optical flow methods, centroid, temporal differences, templates and Spatio-temporal volume. Certain Templates are used for recognizing features by comparison. Certain motion analysis templates are MEI and MHI. A higher version of MEI is CCMEI [4].

Depending on number of cameras used to obtain visual information action recognition techniques can be classified as single view and multiple view ones [17]. Single view methods utilize one camera and multi-view ones utilize multiple cameras set up.

As said before human action recognition comes up with only two main stages (1) A training stage (2) classification stage. The classification is mostly done by SVM or nearest neighborhood classifiers

3. SURVEY ON DIFFERENT PROPOSALS FOR HUMAN ACTIVITY RECOGNITION

A Survey on latest 10 different proposals has been done in the Table.1 It has been noted that the initial stage of extraction is a must in most cases except for depth map extracted from depth cameras or depth sensors. Most of the initialization procedure starts with background subtraction for silhouette extraction. It is also seen that different methods have been adopted to extract the desired features and to reduce the extracted feature dimension for activity representation. An average recognition accuracy rate has also been included based on the authors experiment. The above tabulation shows that SVM classifier is best and most used method for effective activity classification. Abbreviations used in Table 1 listed below

Since most of the above proposals concentrate on feature representation and dimensionality reduction a description of the methods proposed will be discussed. And in the later section discuss about some common classifiers.

3.1 Feature Representation and Dimensionality Reduction

This will discuss the basic steps for action recognition by each proposal in the above survey. Feature Representation and dimensionality reduction for each proposal has been explained below. Feature can be represented using different methods. Contour coding of motion energy image was used to extract global feature from adjusted MEI by square to circular transformation and contour coding [4].SEI is a local motion descriptor defined from silhouettes [2]. Direct silhouette is used to create Spatio-temporal subspace [6]. Silhouette can be used to produce action volumes [18]. STIP method is used for extracting interest points [8, 12, 20]. Trajectories of tracked body joints are used to present time – series representation of activity [11]. MHI is used for event detection [19].Depth Maps is a visual representation for 3D action recognition [20]. These represent features in high dimension.

Dimensionality reduction is done by the following methods. Spare coding and Dictionary learning is used to represent is used in order to extract salient features and present feature in reduced dimensions [7].Adaptive locality Preserving projection is used to reduce spatial dimension from silhouette and NCDAV method is to represent temporal information [6]. Piecewise aggregate approximation (PAA) and Symbolic aggregate approximation (SAX) together reduces dimension of time series representation of tracked body joints [11]. Cluster Discriminative analysis linearly transforms to low dimensional Action vectors from action volumes [18]. Space Time Occupancy pattern is a vector representation from sequence of depth maps and its dimension is reduced by orthogonal class learning (OCL) [20].

Bag Of features [7] and bag of words [8] are descriptors that contain a combined feature representation of Spatio-temporal features. Different local feature descriptors like HOG, HOF, covariance etc. are developed from spatial or temporal features. And hence combined effect of local feature descriptors produce BOF and BOG descriptors. BOF and BOG are more efficient and accurate than other descriptors.

3.2 Classification

Recognition stage mainly has a training stage and classification stage .Based on the survey, it was clear that not much of research is held in the classification stage. The most common classifier that we came cross was SVM and is proved to be best approved classifier. Fig 4 shows SVM is widely used.

SVM is a discriminative classifier and later advanced by Vapnik et.al [21].Even though SVM classifier deals with two category classification problems, multiple SVM used in hierarchical manner can solve multiple classification problems. One algorithm used for SVM in hierarchical manner is SVM-BTA [4].Multiclass SVM is used for training as well as classification [9].SVM classifier has been used by authors from [4, 7-9, 18-20].There are other classifiers used such as K-NN [6], NNC [11], SDSM [12] and HMM [19].

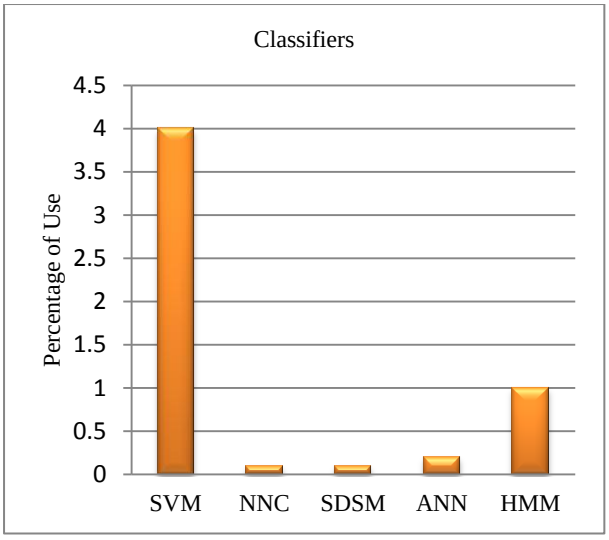

Fig 4: Different Classifiers in Use

4. DATABASES

Different databases are available in-order to test proposed algorithms for human activity recognition. Even though there are different activities available, most of them are created in specified environments. These databases help to test variety of activities up to certain extend. There are different databases available for human activity recognition [10]. Here we shall list out some Database that came across in survey.

4.1 KTH Database

This is a video database created by KTH Royal Institute of Technology in 2004. It is a video database with sequences of human actions and database consists of 6 types of human actions such as boxing, waving, walking jogging, hand waving and hand clapping. Ground truth data i.e. Description of the activity in the scene is provided [10].

4.2 TV Human Interaction Dataset (TVHID)

Visual Geometry group created TV Human Interactions Dataset in 2010. This dataset consists of 300 video clips collected from 20 different TV shows and contain four activities such as handshakes, high fives, hugs and kisses. This dataset mainly provides interaction between two persons [10].

4.3 WEIZMANN datasets

The Weizmann Institute of Science provides two different datasets [10]. They are

Weizmann event-based analysis: This dataset was created during 2001 and was recorded for studying algorithms for clustering and temporal segmentation of videos using statistical measures [10].

Weizmann actions as space-time shapes: This dataset was created and recorded in 2005 for studying new algorithms that improved the human action recognition It contains 10 human actions jumping , walking, running, , skipping , galloping sideways, bending, one-hand waving, two-hands waving, jumping in place and jumping jack [10].

4.4 UCF sports dataset

This has been collected from various sports videos. The actions in this dataset include 8 actions like diving, kicking, horseback riding, golf swinging and running, swinging a baseball bat and lifting. These actions are seen in a wide variety of scenes and hence it is a good database for testing purpose [12].

4.5 I3Dpost dataset

This is a multi-view action recognition database. Contain 8 actions ad two personal interactions. Actions are walk, run, jump forward, and jump in place, bend, sit and fall and wave one hand. Two interactions are hand shaking and pull [18].

4.6 MSR action dataset

This dataset was recorded by using depth acquisition device using infrared light. Contain 20 action types by 10 subjects [20]

Table 1. Survey on different proposals for human activity recognition

	Feature Extraction and Representation						Recognition	Dataset	Accuracy
	EXTRACTING AREA OF IMPORTANCE		HUMAN ACTION INTERPRETATION				METHOD		
	Human Area	Salient Region	Shape Extraction By	Motion Capture By	Template	Spatio-temporal			
[4]	BS	-	Minimum Bounding	Centroid	CCMEI	-	SVM-BTA	Schultd	88.69
[9]	BS	-	-	-	SEI	-	MC-SVM	KTHDB	88.5
[7]	-	Dictionary learning	-	-	-	HOG, Covariance descriptors	BOF, SVM	KTHDB	90.1
[8]	-	STIP	-	-	-	HOG, HOF	BOW, SVM	TVHID	88
[6]	BS	-	-	-	-	NCDAV	LMNN, K-NN	Weizman	98.9
[11]	-	-	-	Trajectories of body joints	-	-	NNC	Weizmann	88.6
[12]	-	STIP	-	-	-	HOG, HOF,CMS	SDSM	UCF sports	86.6
[18]	BS	-	-	-	Action Volume		SVM/ANN	I3DPost	94.44
[19]	BS	-	-	Anthropometric profile	MHI	-	HMM, SVM	ACTIBIO	89
[20]	-	-	-	-	Depth Maps	-	SVM	MSR Action	98.41

5. CONCLUSION

This paper is to provide an overall insight to human activity recognition from feature extraction to human activity recognition and latest methods for analysis that have been adopted. It also tries to generalize the overall activity analysis. Human activity recognition does not impose or stick to any rules or constraints and so the analysis can be done by any technique that suits the purpose of activity recognition. The ideal case of any vision based human activity recognition should be independent of the background of moving targets , noise factors , environment at which human activity is performed, human age and the most important should be view-invariant i.e. independent in camera parameters and its orientation. There are view-invariant techniques adopted like fundamental ratio method [13], multiple camera methods [14-15], 3D gesture recognition [16].

The final stage of HAR, which is the classification of activity recognition, is almost similar. The common procedure adapted is that they have a trained set of data and compare the incoming activity to be recognized with the trained data. It is based on the training set we identify the activity from the input activity source. In order to train we need a quantized set of data/feature representation to identify the activity uniquely. And so the most crucial point in identifying human activity lies in feature extraction and its representation accurately. The more accurate the feature representation the more accurate will be the activity recognition. Most of the above papers focus on dealing with extracting Spatio-temporal features since they provide descriptors based on both space and time combined and provides accurate descriptors [7-8, 12] of features. Mostly adopted methods for HAR classification is by multi-level SVM, HMM or ANN. Even though there are proposals for recognition rate above 95%, most of them have disadvantages of recognizing any particular activity instead of another or depends on recognition in specific environment. On an average till to date we can expect to have atleast 90%

for different activity recognition. This is not sufficient since there are chances of errors. Aiming recognition rate of minimum 98% in any environment is a must for error free recognition. Wide variety of applications and aim for error free recognition led HAR to hold space still as a research area in real-world.

6. ACKNOWLEDGMENTS

We would like to express my gratitude to our Alma mater VJEC and grateful to our principal Dr. Benny Joseph as he is leading light of the institution. We are indebted to our HOD Mrs. Roshini T. V for her encouragement. Grace our gratitude to all faculties of department of ECE who helped us directly or indirectly towards the successful completion of this paper.

7. REFERENCES

[1] Salah Althloothi, Mohammed H. Mahoor, Xiao Zhang, and Richard M.Voyles "Human activity recognition using multi-features and multiple kernel learning," Pattern recognition, vol. 47, pp. 1800-1812, Dec 2013.

[2] Attila Reiss, Gustaf Hendeby ,Dider Stricker, "A Competitive approach For Human activity recognition on smart phones", in Google Scholar, pp.455–460.

[3] Ankun Agarwal , Bill Trigis, "Recovering 3D Human Pose from Monocular images ," IEEE transactions on pattern analysis and Machine intelligence, vol.28 ,pp.44-58,Jan 2006

[4] Huimin Qian, Yaobin Mao, Wenbo Xiang, Zhiquan Wang, "Recognition of human activities using SVM multi-class classifier" Pattern Recognition Letters,vol 31, pp.100-111,Sep 2009 .

[5] Ju Han, Bhanu B, "Human Activity Recognition in Thermal Infrared Imagery ," IEEE Computer Society Conference on computer vision and Pattern Recognition,pp.17,Jun 2005

[6] Chien –Chung Tseng,Ju-chin Chen, Ching-Hsien Fang, Jenn-Jier James Lien ,"Human action recognition based on graph embedded spatio-temporal subspace,"Pattern recognition,vol.45,pp.3611-3624,April 2012

[7] Guruprasad Somasundaram , Anoop Cheriyan,Vassilios Morellas, Nikolaos Papanikolopoulos, "Action Recognition using Global Spatio-Temporal Features Derived from Sparse Representations" , Jan 2014

[8] Manuel J. Marin-Jimenez, Enrique Yeguas, Nicolas Perez de la Blanca, "Exploring STIP-based models for recognizing human interactions in TV videos", Pattern Recognition Letters,vol.34,pp.1819-1828,2013

[9] Mohiddin Ahmad, Seong – Whang Lee,"Variable silhouette energy image representations for recognizing human activities," Image and Vision computing , vol.28 , pp.814-824,2010

[10] Jose M.Chaquet,Enrique J.Carmona, Antonio Fernandez Caballero,"A survey of video datasets for human action and activity recognition ," Computer Vision and Image Understanding,vol.117,pp.633-659,Feb.2013

[11] Imran N.Junejo,Zaher Al Aghbari , "Using SAX representation for human action recognition," J.Vis.Commun.Image R.,vol. 23 ,pp.853-861,2012

[12] Harong Wang , Chunfeng Yuan , Weiming Hu , Changyin Sun, "Supervised class-specific dictionary learning for sparse modeling in action recognition ," Pattern Recognition, vol.45, pp.3902-3911,Apr2012

[13] Nazim Ashraf,Yuping Shen , Xiaochun Cao, Hassan Forrosh "View invariant action recognition using weighted fundamental ratios," Computer Vision and Image Understanding , vol.117,pp.587-602,Feb2013

[14] M. Ahmad, S. Lee, "HMM-based human action recognition using multiview image sequences", ICPR vol.1,pp.263–266, 2006

[15] F. Cuzzolin "Using bilinear models for view-invariant action and identity recognition", Proceedings of CVPR, pp.1701–1708,2006

[16] L. Campbell, D. Becker, A. Azarbayejani, A. Bobick, A. Pentland, "Invariant features for 3-d gesture recognition," FG pp.157–16,199

[17] P. Turaga, R. Chellappa, V. Subrahmanian, O. Udrea, "Machine recognition of human activities: a survey," IEEE Transactions on Circuits and Systems for Video Technology vol.18,pp.1473–1488,Nov2008

[18] Alexandros Iosifidis , Anastasios Tefas, Ioannis Pitas, "Multi-view action recognition based action volumes, fuzzy distance and cluster discriminate analysis" Signal Processing 93,pp.1445-1457,2013

[19] Anastasios Drossou,Dimosthenis loannidis,Konstantios Moustakas,Dimitrios Tzovaras, "Spatitemporal analysis of human activities for biometric authentication," Computer Vision and Image Understanding 116, pp.411-421, 2012

[20] Antonio W. Vieira,Erickson R.Nascimento, Gabriel L. Olivera, Zicheng Liu, Mario F.M. Campos, "On the improvement of human action recognition from depth map sequences using Space-Time Occupancy Patterns" Pattern Recognition Letters 36,pp.221-227,2014

[21] Vapnik, V., Statistical Learning Theory. John Wiley and Sons Inc., New York. Veeraraghavan, A., Roy-Chowdhury, A.K., Chellappa, R. "Matching shape sequences in video with applications in human movement analysis". IEEE Trans. Pattern Anal. Machine Intell., vol.27, pp.1896–190, 2005