Automatic Recognition of Human Activities under Variable Lighting

Jaime R. Ruiz[1], Leopoldo Altamirano[2], Eduardo F. Morales[3], Adrián León[4], and Jesús A. González[5]
{jrruiz[1] ,robles[2] ,emorales[3] ,enthe[4] ,jagonzalez[5] }@ccc.inaoep.mx

Department of Computer Science
National Institute of Astrophysics, Optics, and Electronics
Luis Enrique Erro #1, Sta. Maria Tonantzintla,C.P. 72840
Puebla, Mexico.
Tel.:+52 222 266 3100, ext, 8303; Fax: +52 222 266 3152.

Abstract. The recognition of activities plays an important role as part of the analysis of human behavior in video sequences. It is desirable that monitoring systems may accomplish their task in conditions different to the training ones. A novel method is proposed for activity recognition on variable lighting. The method starts with an automatic segmentation procedure to locate the person. It takes the advantage of Harris and Harris-Laplace operators of capturing information in spite of extreme changing lighting to locate corners along the human body. Corners are followed through the images to generate a set of trajectories that represent the behavior of the human. The method shows its effectiveness recognizing behaviors by a comparison procedure based on dynamic time warping, and also working well with examples of activities under different lighting.

1 Introduction

The analysis of human behavior has taken great importance in modern surveillance systems, due to its application for video analysis, elderly care, video retrieval, among others.

In the last decade, the demand for systems capable of interpret human behavior in video sequences and capable of operating correctly under variable lighting conditions has been an unsolved challenge.

This capability depends directly on the performance of an algorithm to determine the location of a person on the scene and to follow it through the next images in the video sequence. This tracking information is not enough for determining what the person does at the place. If the algorithm can obtain consistent information, we need after that a learning phase that defines how the information will be represented, and how a model will be constructed to represent the individuals' activities, and in later steps to identify similar activities.

A big effort has been dedicated, and a lot of works have been proposed to accomplish this work. Some of the methods involve to determine activities based

in the human body's form, principally based on silhouettes, to determine what the person did at the place of analysis [15,8,1], or if the activity is normal or not, [12,6]. These approaches do not work on scenes with variable lighting.

Other works like [16,2,5] are able to recognize activities according to the person's movements and consider gradual changes in environment lighting. However, these approaches consider the tracked person as a whole region. As a consequence of this simplification, they cannot distinguish activities where articulations as individuals are involved. Considering both trends, our work can do the recognition of activities based on the human body articulations taking into account a scene with variable illumination.

The method use local feature operators as a tool to segment and follow the person inside a scene. From these operators, we calculate space-time information derived from the tracking of interest points. After that the algorithm builds an activity model using the tracked points in a compact form. We choose b-splines to represent the trajectories over the scene. Once the models have been constructed, the activities are evaluated with a test set of activity sequences taken in different lighting conditions. The results show that the method can be used to recognize activities in this kind of environments.

The organization of the paper is as follows. In the section 2, the method of feature extraction and the segmentation used are described. In the section 3, we detail the representation of the trajectories obtained based on b-splines. The generation of the model and how to recognize the activities are explained in section 4. Experiments and results are presented in section 5, and conclusions are exposed in section 6.

2 Feature extraction method

2.1 Features

Follow a person under changes of illumination in a scene represents a challenge for tracking algorithms. In spite of this, there are algorithms that can be used for this purpose. These approaches works by taking into account the spatial information of the object of interest in an initialization step. Then, in subsequent frames, the algorithms calculates the new object position in the scene using data from the previous frame.

Finding out the location in the scene of the tracked person is not a hard task if we need only to know where the person is situated [16,2]. However, if we need to make an analysis based on the human body parts is essential not only to know its position in the environment but also to determine the area that it occupies in the scene with the purpose of obtaining information from different parts of its body. With these data it is possible to make an analysis using the pose of the person.

Taking into account the idea explained above,and knowing the complexity involved to do this on variable lighting conditions. We perform an exhaustive

evaluation of some interest points algorithms reported in the literature. We evaluate SIFT, Hessian-Laplace and Harris-Laplace algorithms under different degrees of illumination; the results have been showed in the table 1. According with the results, we have proposed the use of Harris-Laplace detector [11] to obtain a set of points located on several parts of the human body, and capture motion information of these. This can be seen in figure 1. An analysis about the detector Harris-Laplace allows us to know that some corners have been removed by the operator because they do not pass the selection criteria for multiples scales —[11]— therefore there is information not considered that may be helpful. As a result, to compensate this loss, we propose to use the original Harris detector [7] for getting a greater number of points on the region of interest.

	Five Activities		
	SIFT	Harris-Laplace	Hessian-Laplace
Morning	168.6	274.4	209.6
Evening	34.8	175.4	53.6
Night	12.4	109.2	16.2

Table 1: Points number average calculated for each detector under three lighting conditions.

Fig. 1: Example of Harris-Laplace points calculated with the proposed method.

2.2 Segmentation

For exploiting the robustness of the Harris points to extreme lighting, we use them in the segmentation procedure to locate the person in scene. The method starts by applying the detectors of Harris and Harris-Laplace on the first frame, to obtain a set of points.

After that, we examine the behavior of this collection of points during the next nine frames, with the purpose of finding regions with high movement.

To do this, we work with tracking methods based on predictions capable of operating with variable lighting. In this direction, the algorithm initially proposed by Lucas and Kanade in [10], which was later fully developed by Tomasi and Kanade in [14], is used as our tracking module.

It allows the tracking of multiple points in a sequence of images as shown in [14]. There are several variants of the Kanade-Lukas-Tomasi tracker (KLT), and we use in this work its pyramidal implementation.

Once the points have been tracked, we need to determinate the regions of interest. In this case we define a threshold α to identify zones with high motion. During the frames, the algorithm accumulates the displacement calculated by the KLT tracker for each point obtained in the initial stage. If the points displacement is higher or equal to the threshold α after ten frames, then they are considered in the next phase as moving objects in the scene.

Next, to situate the person on the scene and discriminate it of others objects in movement, we employ the basic idea of the traditional segmentation methods. We suppose a minimum size β of the person as a new threshold to eliminate regions below it. After that, we can know where the person is. The general idea of segmentation procedure can be seen in the figure 2.

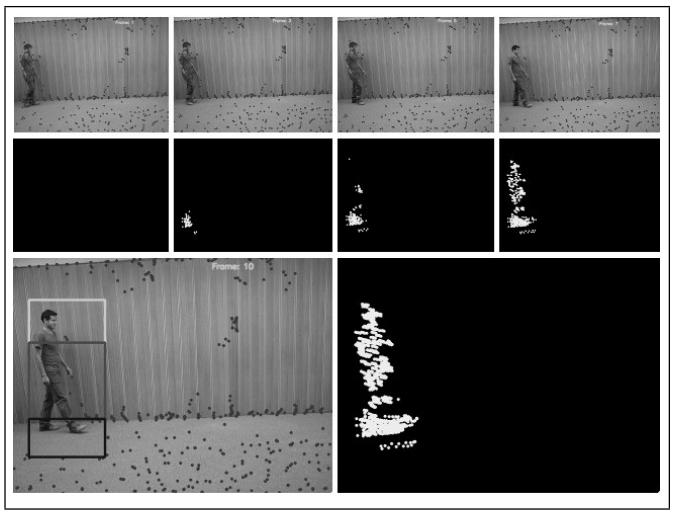


Fig. 2: Example of the segmentation procedure. First row: Motion of the points over the 1,3,5,7 frames. Second row: Points higher or equal to threshold α for 1,3,5,7 frames. Third row, left: Region of interest after ten frames (three rectangles box) and right: set of points higher or equal to threshold α after ten frames.

2.3 Tracking

Once the person position and size are determined, the next is to determine which strategy is useful to capture the motion of the body of the person being tested.

On the same line, the KLT algorithm uses Harris points calculated at the initial step to find the new location of these points on the next frames, after a period of time. This results in a set of trajectories that are taken as motion information of the human body parts. In this way, the trajectories can reflect what the person performs in the scene. The behavior of the points along the sequences is depicted in figure 3.

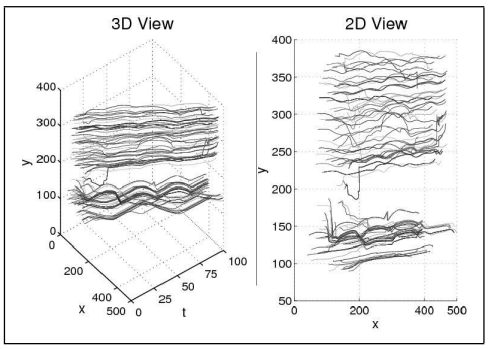


Fig. 3: Example of Harris points trajectories in the "Walk" activity. Left: spatio-temporal view for tracked points with KLT algorithm, and right: 2D projection to x and y components for tracked points.

Notice that the number of trajectories obtained from this approach may contain redundant information of different areas of the body. For this reason, we established three major regions of analysis on the size of the person, top, middle and bottom. These regions are defined at the beginning of the algorithm.

Defined the three regions, we need to determine the points that are within the limits of these to create three sets of points, $Top = \{P_1, \ldots, P_L\}$, $Middle = \{P_{L+1}, \ldots, P_M\}$ and $Bottom = \{P_{M+1}, \ldots, P_N\}$. Where N is the total of Harris points calculated when the algorithm starts.

Subsequently, the algorithm creates a central point. It can be viewed like an average of all points in each one of the sets, according to each new prediction of the tracking algorithm until the time T in which the activity ends. It is calculated through the following expressions:

$$topCP(x_t, y_t) = (\max_{i=1,\ldots,L} x_i, \sum_{i=1}^{L} y_i/L) \tag{1}$$

$$middleCP(x_t, y_t) = (\max_{i=L+1,\ldots,M} x_i, \sum_{i=L+1}^{M} y_i/(M-L)) \tag{2}$$

$$bottomCP(x_t, y_t) = (\max_{i=M+1,\ldots,N} x_i, \sum_{i=M+1}^{N} y_i/(N-M)) \tag{3}$$

Note that in the expressions $(1, 2, 3)$. $t = 1, \ldots, T$, the position x_t is the coordinate farther in the x axis. This is the direction in which the person is moving. This is done to preserve as much detail in the trajectory generation as possible. The y_t position is the simple average of the y-coordinates of the points within each set.

Once a sequence of T frames was analyzed, we obtain three resulting curves that encode the trajectories generated by the tracking algorithm to this instant, see figure 4. In this way, the information of the different parts of the body can be processed. Based on this information we can determine the activity that the person is carrying on.

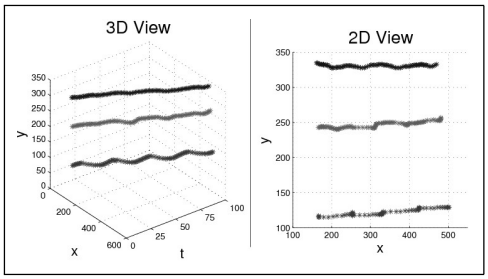


Fig. 4: Example of the central points for "Walk" activity after $T = 100$ frames. Left: spatio-temporal view for tracked points with KLT algorithm, and right: 2D projection to x and y components for tracked points.

3 Representation of trajectories

As shown in figure 4 there are variations in the resulting paths due to errors in the predictions of the tracking algorithm as a consequence of changes in lighting. Therefore, in order to have consistent information and somehow eliminating these variations in the curves produced, we propose the use of uniform cubic b-spline curves [3]. In this way we can generate a compact model based on curves, avoiding the storing of all information corresponding to the trajectories.

According to its definition, a b-spline can approximate and smooth a collection of points with the combination of a set of size P of basic polynomial functions and a collection of size P of coefficients. The b-spline provides an adjustment for the original curves.

In our study, each of the curves is parameterize by time. The algorithm generates a b-spline for all X components with respect to t and a b-spline for

Y with respect to t. At the end, the information is combined to generate the resulting curve that encodes the trajectory of each of the zones established for the analysis. As an uniform b-spline is used, we need a knot vector -[3]- evenly distributed to make the adjustment.

The procedure can be summarized as follows: The central points of each area identified -top, middle and lower- are stored in the curves $TopCurve$, $MiddleCurve$ and $BottomCurve$, respectively.

Each of these curves is then approximated and smoothed with b-spline, as previously established. In this way models are built for the top, middle and bottom zones of the person for each of the activities by eliminating the effects of different lighting conditions on the scene. The resulting curves can be seen in figure 5a.

Once patterns of activity are generated and the effects of lighting changes are eliminated, see figure 5, we must define how a new activity will be identified with respect to the activities stored.

4 Recognition of activities

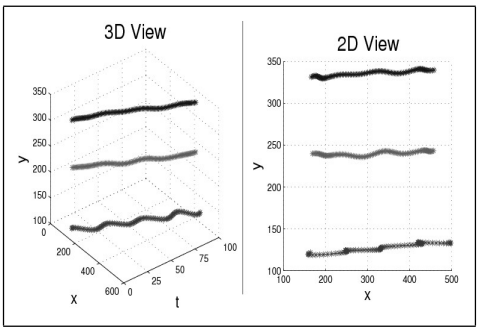
Each activity consists of a set of three curves, resulting from the approximation to b-spline of the components X y Y of each curve. Due to the nature of the people's activities, an activity can be re-executed at diverse speeds either by the same person or by a different person. Therefore we must use a method that allows us to compare the shape of two curves which have different proportions. It should determine a degree of similarity between them. In this case, we use the dynamic time warping algorithm (DTW) [13] to do this comparison.

The DTW was developed for comparing two time series of different lengths and find out the similarity grade between them [13]. This technique is widely used in mathematical field [4,9]. Using these algorithms is possible to find similarities between signatures and person activities based on sensors.

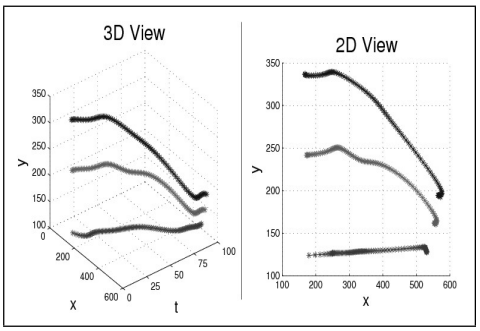
It is worth to mention that it is necessary to follow the same procedure that was used for the generation of the stored models in the moment of the comparison of an unknown activity with the stored activities models. This means that we need to generate a model with three curves that represents the new behavior. For a better adjustment in the comparison step, the models and the activities to be evaluated are moved to one common point that we will call origin.

Then, we compare one by one the curves corresponding to each zone of the person. We compare the top curve of the activity that we attempt to recognize with each of the top curves of the activities that were stored as models using DTW. With this action, we obtain a degree of similarity between them. The procedure is the same for middle and lower curves.

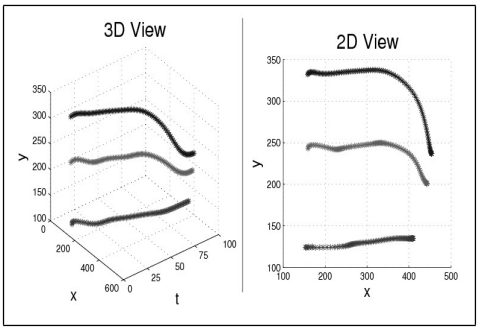
At the end of comparison of the zones, the recognized activity will be the activity with the highest fit, i.e., activity with the highest average measure of similarity of the three zones. Thus the method can recognize activities through DTW as a measure of similarity between the models.



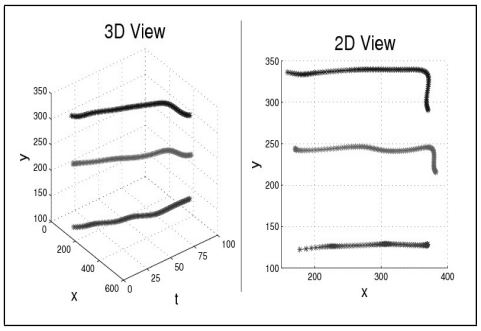
(a) Walk



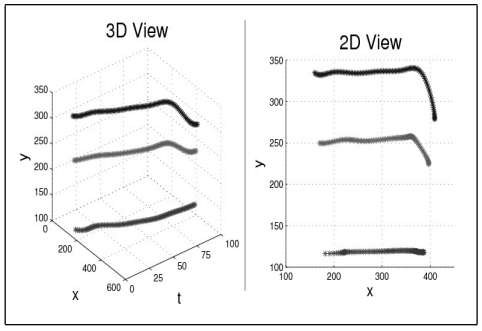
(b) Fall



(c) Crouch



(d) Sitting Front

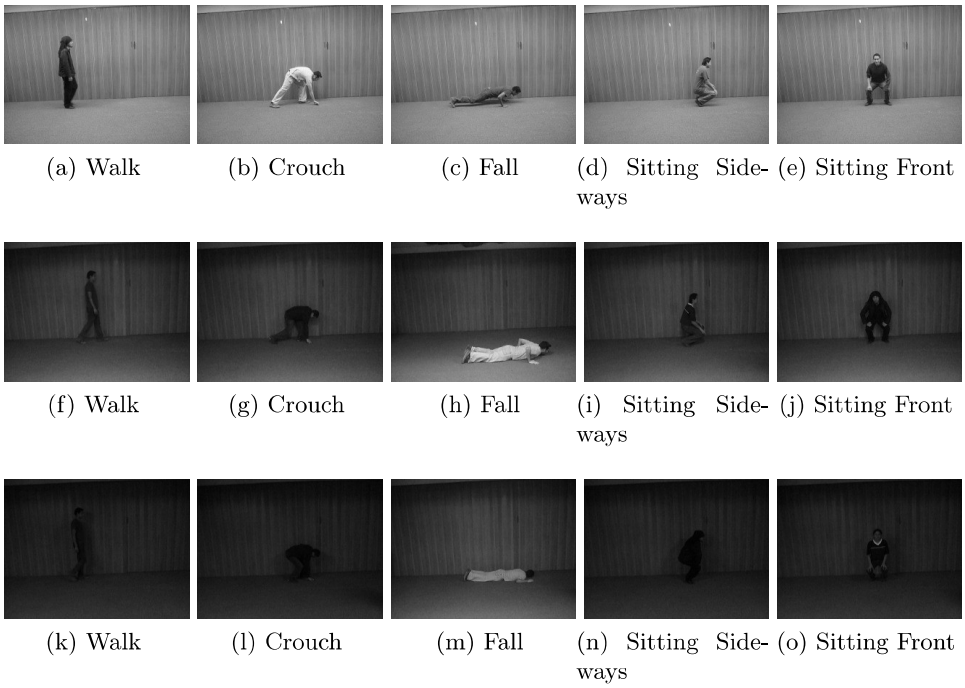


(e) Sitting Sideways

Fig. 5: Activity models generated by the b-splines curves. For each image, the left side shows the spatio-temporal view for b-spline curves, and right side shows the 2D projection to x, y axis for each b-spline curve.

5 Experiments and results

This section details the experiments performed to test the proposed method. In our case, five people were employed with different anatomy and clothing as objects of study. Five basic activities were carried on in a static background and a fixed camera with front sight to the scene and three different lighting changes to test the method were tagged as, Morning, Evening and Night, as can be seen in figure 6. Finally, every single person reproduced the five activities under different lighting conditions to generate a total of 15 analysis sequences for each of the activities.



(a) Walk (b) Crouch (c) Fall (d) Sitting Sideways (e) Sitting Front

(f) Walk (g) Crouch (h) Fall (i) Sitting Sideways (j) Sitting Front

(k) Walk (l) Crouch (m) Fall (n) Sitting Sideways (o) Sitting Front

Fig. 6: Examples of sequences used to test the proposed method. First row: Morning. Second row: Evening. Third row: Night.

The training phase was done using the morning sequence, and the model base was generated by showing a single sequence of each activity. The method was tested using 75 sequences of activities that included sequences of the morning, evening and night. The thresholds considered in this phase are $\alpha = 50\,pixels$ and $\beta = 45,000\,pixels^2(150\ pixels\ X\ 300\ pixels)$.

For the generation of models, interpolations and smoothing by b-spline were made with a knot vector uniformly distributed according to the interval in which

each curve was evaluated. We used 10 control points and therefore, 10 basic functions, and 10 coefficients were used to adjust the curves that were analyzed.

For the purpose of comparison we use other approach. A human operator localize and establish the area of the person manually in the beginning of the method.

The results of the proposed method are summarized in the confusion matrix in table 2. The results of the method assisted by a human operator are concentrated in the confusion matrix in table 3, both results show the average score of repeating the experiment five times.

At the end a recognition rate of 88% in average was reached for all activities for the proposed method in comparison with a recognition rate of 89.33% for the manual method.

	Walk	Crouch	Fall	SF	SS
Walk	15				
Crouch		14			1
Fall		1	14		
SF				14	1
SS		5		1	9

Table 2: Confusion matrix with the results from the recognition of activities using the automatic proposed method.

	Walk	Crouch	Fall	SF	SS
Walk	15				
Crouch		14			1
Fall		1	14		
SF				14	1
SS		3		2	10

Table 3: Confusion matrix with the results from the recognition of activities using the method assisted by a human operator.

6 Discussion

The percentage of correct classification of the results presented in table 2 may seem low compared with other approaches reported in the literature. However, the proposed method is evaluated in a scenario with different levels of illumination. Furthermore, compared to other jobs with uncontrolled lighting, this project, in very similar activities such as “Crouch” and “Sit”, had good results. But, the activity “Sitting sideways” was confused with the activities of “Crouch”

and "Sitting front" because sometimes it was impossible to obtain a set of curves to make a clearer distinction between activities. By contrast, in "Walk" activity, the method has no problems extracting the curves to make the distinction between "Walk" and other activities. Comparing the method proposed with the manual approach, even though we get a better percentage in the correct classification with the human assisted method, the results show that only one example correctly classified in the "Sitting Sideways" activity makes the difference.

7 Conclusions

The proposed method is useful for recognition of activities under different lighting conditions, using a simple technique for comparison.

The procedure in the same way, gives evidence that Harris operator can locate important body parts for the analysis of behavior, in our case, the head, torso, legs and feet, and these features can be detected even though there is a change in lighting conditions. The algorithm is also able to recognize activities based on the human body that are similar and can be easily confused between them.

As future work, we plan to test the method performance with abrupt changes in lighting when the activities are performed in outdoor scenarios.

Acknowledgment

The research reported in this paper was supported by the National Council of Science and Technology of Mexico (CONACYT) scholarship No. 40427.

References

1. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 288–303 (2010)
2. Arasanz, P.B.: Modeling Human Behavior for Image Sequence Understanding and Generation. Ph.D. thesis, Universidad Autónoma de Barcelona, España (2009)
3. Deboor, C.: A Practical Guide to Splines. Springer-Verlag Berlin and Heidelberg GmbH & Co. K (dec 1978), http://www.worldcat.org/isbn/3540903569
4. Efrat, A., Fan, Q., Venkatasubramanian, S.: Curve matching, time warping, and light fields: New algorithms for computing similarity between curves. J. Math. Imaging Vis. 27, 203–216 (April 2007), http://portal.acm.org/citation.cfm?id=1265122.1265128
5. Fernández Tena, C., Baiget, P., Roca, X., González, J.: Natural language descriptions of human behavior from video sequences. In: KI 2007: Advances in Artificial Intelligence, pp. 279–292 (2007)
6. Goya, K., Zhang, X., Kitayama, K., Nagayama, I.: A method for automatic detection of crimes for public security by using motion analysis. In: Proceedings of the 2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. pp. 736–741. IIH-MSP '09, IEEE Computer Society, Washington, DC, USA (2009), http://dx.doi.org/10.1109/IIH-MSP.2009.264

7. Harris, C., Stephens, M.: A combined corner and edge detection. In: Proceedings of The Fourth Alvey Vision Conference. pp. 147–151 (1988)
8. Lao, W., Han, J.: Flexible human behavior analysis framework for video surveillance applications. International Journal of Digital Multimedia Broadcasting 2010, 1–10 (2010)
9. Liu, J., Wang, Z., Zhong, L., Wickramasuriya, J., Vasudevan, V.: uWave: Accelerometer-based personalized gesture recognition and its applications. Pervasive Computing and Communications, IEEE International Conference on pp. 1–9 (2009), http://dx.doi.org/10.1109/PERCOM.2009.4912759
10. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI81. pp. 674–679 (1981), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.2019
11. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark. pp. 128–142. Springer (2002), http://perception.inrialpes.fr/Publications/2002/MS02, copenhagen
12. Nater, F., Grabner, H., Gool, L.V.: Exploiting simple hierarchies for unsupervised human behavior analysis. Computer and Robot Vision (CRV2010) (2010)
13. Sakoe, H.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 26, 43–49 (1978)
14. Tomasi, C., Kanade, T.: Detection and tracking of point features. Tech. rep., International Journal of Computer Vision (1991)
15. Wang, Y., Mori, G.: Human action recognition by semilatent topic models. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(10), 1762–1774 (2009), http://www.ncbi.nlm.nih.gov/pubmed/19696448
16. Zhou, Z., Chen, X., Chung, Y.C., He, Z., Han, T.X., Keller, J.M.: Activity analysis, summarization, and visualization for indoor human activity monitoring. Circuits and Systems for Video Technology, IEEE Transactions on 18(11), 1489–1498 (2008), http://dx.doi.org/10.1109/TCSVT.2008.2005612