A Framework for Image Enhancement via Contextual Information and Epitome-based Representation

Chee Seng Chan, Honghai Liu, Tsz Ming James Hui, David J. Brown and [1]Marzuki Khalid

Computer Intelligence & Applications Group,
Faculty of Technologies,
University of Portsmouth,
Portsmouth, PO1 3HE,
UNITED KINGDOM.
{cheeseng.chan;honghai.liu;james.hui;
david.j.brown}@port.ac.uk

[1]Centre of Artificial Intelligence and Robotics
(CAIRO), Universiti Teknologi Malaysia,
Jalan Semarak, 54100 Kuala Lumpur,
MALAYSIA.
marzuki@utmkl.utm.my

Abstract

Image enhancement in our understanding includes quality improvements and understanding improvement of a digital image or video. Our study focuses on the latter; this paper proposes a framework for image understanding using epitomes and contextual information in human-motion involved object recognition systems. The proposed has been raised when considering image enhancement with limited computational resources. We demonstrate the proposed with an example of recognizing a target object, i.e., a watch, on a walking man. First, we employ unsupervised learning to train the video in terms of epitomes, and then the classification of the watch is formulated as a search problem of finding the target pixels in the epitomes. Though the quality of the regenerated object is relatively low, the watch can be recognized with the aid of recent results from human motion analysis.

1. Introduction

Over the past years, human motion analysis and object recognition has been an active research area in the computer vision research community. This interest is motivated by the availability of low cost commercial product for capturing video which opens up a wide spectrum of topics such as in surveillance area and control area (see [22,39] for a survey).

Different approaches have been proposed for human motion analysis [see [33] for a survey]. Two typical approaches to human body tracking are depending on whether a priori shape models are used. In both model based and nonmodal based approaches, the representation of human body has evolved from stick figure [19] to 2D contour [2,33] to 3D volumes [9,15] as complexity of the model arises. Model based approaches such as [30,19,15] basically require manual initialization in the first frame. Example is [31,24] where a cardboard model [7] were used as human model.

Approaches that without pre-defined human models, such as in [33] is heavily dependent on heuristic assumptions which impose constraints on feature correspondence and decreasing in search space. Example is [18] where it used a statistical approach which is similar to taken in [31,24] but it attempts to make the background subtraction more robust to environmental dynamic by incorporating noise measurement in the system.

Conventional human motion analysis on video streams was proposed in [31,24,18,8] but most of these systems for human tracking are based on Kalman filter. As pointed in [25,26], they are inadequate because it is based on Gaussian densities which is unimodal and cannot represent simultaneous alternative hypotheses. So Isard and Blake [26] proposed CONDENSATION - a stochastic framework for highly robust tracking. Then Bayesian approach is proposed to detect and track human motions. It showed promising result as it provides a principal probabilistic framework to combine multiple cues and introduce a priori knowledge or constraints related to the class of object to detect and track in the scene [17]. The key point in this approach is to provide an appropriate statistical characterization of the entities of interest and background. Recent work in this area included [1,3,13,27].

Recently, particle filters (a.k.a. CONDENSATION [26], sequential Monte Carlo) has been widely used in human body tracking [9,5,12,29] where it based on sampling approximation and likelihood computation. Due to this, it is capable to handle pose ambiguities subject to motion singularities and occlusions [20]. But there are some problems associated in this. As stated in [6], various method has been proposed on improving the efficiency of using particle filter for human tracking such as covariance scaled sampling [12], simulated annealing [20] , partitioned sampling [5] and hybrid Monte Carlo [35]. Other related works included [36,29]. But large computation complexity still remains an important issue for using this technique in human tracking. Lee in [6] introduced auxiliary measurement to reduce the computation complexity of particle filter. Both motion cue and body detection are included in the auxiliary measurement where motion cue is used to improved sampling distribution of providing better estimation of the current of body parts, while body detection is used to updating subsets of the state parameter, within the particle filtering scheme.

Research on multiple human detection and tracking was proposed in [21,22], but all these focused on low occluded scene. So, Zhao and Nevatia in [4]

proposed Markov Chain Monte Carlo (MCMC) method to segment each individual human in a complex scene from a static camera.

On the other hand, object recognition and classification has made significant process recently since its beginning (see [39] for a survey). It is widely used in the machine vision industry for the purposes of inspection, registration, and manipulation. The challenge of object recognition in cluttered real world scenes is it's require local image features that are unaffected by nearby clutter or partial occlusion. The features must be at least partially invariant to illumination, 3D projective transform, and common object variations.

Traditionally, contour and shape based methods are regarded most adequate for handling the generalization requirements needed for object recognition while appearance based methods such as eigenspace matching [32],color histogram [10] and receptive field histogram [40] have been successful in object identification and detection scenarios.

In [40], a technique to determine the identity of an object in a scene using multidimensional histograms of a vector of local neighborhood operators is proposed. This technique can be used to determine the most probable object, independent of the object's position, image plane orientation and scale. Then Schiele and Crowley in [38], a probabilistic object recognition technique which does not require correspondence matching of images is proposed. This method is an extension of [40].

Then Lowe [23] developed an object recognition system that used a new class of local image features. The features are invariant to image scaling, translation and rotation. It also partially invariant to illumination changes and affine or 3D projection.

In this paper, the authors proposed a novel method for human-motion involved object recognition systems which uses contextual information and epitome [16] technique. The authors formulated the image enhancement in this system as a search problem of finding the target pixels in the epitome. Then, the result is combined with contextual information from human motion analysis to classify an object. The basic module of this proposed framework is illustrated in Figure 2, and will be discussed in Section 4.

2. Video Epitome

2.1 Epitome Model

Epitome introduced by Jojic et al [37] is an appearance model where it is a condensed version of an image containing the essence of the textural and shape properties of the image. Then, Cheung et al in [16] extended 2D epitome introduced in [37] through time to form a space-time epitome.

2.2 Learning Video Epitome

The task to learn epitome entails minimizing of the Helmholtz free energy cost function:

$$F = \sum_{s} \sum_{T_s} \int_{C_s} q\{(T_s, C_s)\} \log \frac{q(\{T_s, C_s\})}{p(v, \{C_s, T_s\})} \quad (1)$$

It turns out that the free energy [34] lower bounds the log-likelihood of the input video. The lower bound can be subsequently simplified by using posterior distribution. So, by choosing an appropriate Q form, two goals can be achieved (see [16] for detail).

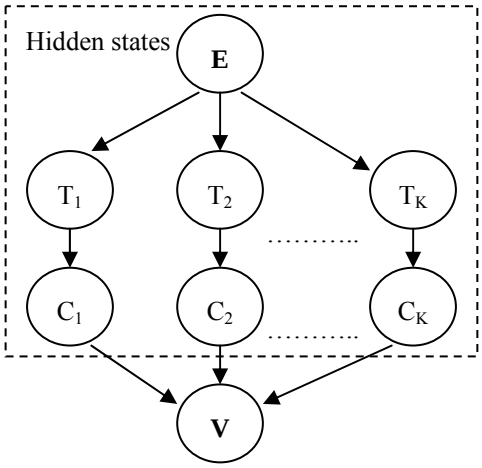


Fig. 1: Generative Model of Epitome of a Video

The free energy obtained using the above q-distribution leads to a traceable iterative learning algorithm. Maximizing in this way, the lower bound of negative Helmholtz free-energy results in standard EM algorithm. So, four updates were achieved.

$$v_{x,y,t} \rightarrow \frac{\frac{v_{x,y,t}}{\sigma^2_{x,y,t}} + \sum_{s,k:s(k)=x,y,t} \sum_{T_s} q(T_s) \frac{\mu_{T_{s(k)}}}{\Phi_{T_{s(k)}}}}{\frac{1}{\sigma^2_{x,y,t}} + \sum_{s,k:s(k)=x,y,t} \sum_{T_s} q(T_s) \frac{1}{\Phi_{T_{s(k)}}}} \quad (2)$$

$$q(T_s) \rightarrow \frac{p(T_s) e_{T_s}(v_s)}{\sum_T p(T_s) e_{T_s}(v_s)} \quad (3)$$

$$\mu_{x_e,y_e,t_e} \rightarrow \frac{\sum_{T,k:T(k)=(x_e,y_e,t_e)} \sum_s q(T_s = T) v_{s(k)}}{\sum_{T,k:T(k)=(x_e,y_e,t_e)} \sum_s q(T_s = T)} \quad (4)$$

$$\Phi_{x_e,y_e,t_e} \rightarrow \frac{\sum_{T,kT(k)=(x_e,y_e,t_e)} \sum_s q(T_s=T)(v_{s(k)} - \mu_{x_e,y_e,t_e})^2}{\sum_{T,kT(k)=(x_e,y_e,t_e)} \sum_s q(T_s=T)} \quad (5)$$

To learn an epitome from video, four updates above are iterated until convergence. If an epitome is given, then video reconstruction can be achieved by simply iterated using (2) and (3) until convergence (see [16] for details).

3. Modeling Human Body

3.1 Visual Tracking

Here, the authors used method proposed in [11] for tracking human body in video sequence where a human skeleton is modeled as a kinematic chain of rigid bodies which undergo a transformation of rigid motion and shape variations. Tracking is performed by minimizing the energy function:

$$E=\sum_{i=1}^{n} f \frac{\sum_{l=1}^{L} \int_{\Omega_l} (\mathrm{I}_i(g_l(p,\Theta_i))-M_l(p))^2 W_l(p) V_l(p,\Theta_i)\,dp}{\sum_{l=1}^{L} \int_{\Omega_l} W_l(p) V_l(p,\Theta_i)\,dp} \quad (6)$$

using a gradient descent scheme with respect to unknown position and shape of the body parts (see [11] for details).

$$\frac{\partial E}{\partial \Theta_i} \rightarrow \frac{1}{A}\Big(\sum_{l=1}^{L} \int_{\Omega_l} \Delta I_i(p,\Theta_i,l)\big(\nabla \boldsymbol{I}_i^{T}(g_l(p,\Theta_i)\frac{\partial g_l}{\partial \Theta_i}\big) V_l(\mathrm{X},\Theta_i) W_l(p)\,dp + \sum_{l=1}^{L} \int_{\Omega_l} \Delta \boldsymbol{I}_i(p,\Theta_i,l)^2 \frac{\partial V_l}{\partial \Theta_i}(\mathrm{X},\Theta_i) W_l\,dp - E_i(\Theta_i,W)\Big) \sum_{l=1}^{L} \int_{\Omega_l} \frac{\partial V_l}{\partial \Theta_i}(p,\Theta_i) W_l(p)\,dp\Big) \quad (7)$$

$$\nabla W_l E(p_k) \rightarrow \sum_i \delta(d_l(p_k)) \frac{V_l(p_k,\Theta_i)}{A} (\Delta I_i(p_k,\Theta_i,l)^2 - E_i(\Theta_i,W)) \quad (8)$$

4. Method

4.1 Assumption

In this work, the authors make the following assumptions: (1) Human tracking and motion analysis results are achieved using [11]; (2) There is only an object in the video sequence; (3) A watch is the target object in the video; (4) Object classified in the human motion analysis is a watch.

4.2 System Overview

The system proposed in this paper consists of two stages as outlined in Figure 2. In the first stage, a moving human is detected and tracked using a kinematic chain model with 5 links (Figure 3) [11]. The model is manually initialized in the first frame of the input video sequence. Then once this initialization step is completed, the system performs the tracking task by minimizing (6), with respect to Θ and W. A preliminary guess of the object class which based on this result is performed latter.

In second stage, the input video, $I_{x,y,t}$ is transformed into a video epitome, $V_{x,y,t}$ using technique described in [16] where $V_{x,y,t} << I_{x,y,t}$. Here, instead of using traditional methods such as segmentation or edge detector, the target in this system is extracted using video epitome. It is done by locating their distributed location in the epitome. The detected target in each epitome is used to perform image enhancement in latter stage by using the information in object detection to extract the appropriate pixel from video epitome, ν_{x_e,y_e,t_e} for multi-frames reconstruction. Next, the shape property of the object is passed to the system for comparison with result from human motion analysis. In this paper, the authors assumed that a watch was classified by human motion analysis.

4.3 Epitome Analysis

In this paper, the authors focused on the problem of extracting an object in the video epitome (Figure 4). The ultimate goal is to build a system that, if properly configured, can reliably detect and extract a target position in the video epitome for image enhancement. This reconstructed object will then compare with result from human motion analysis for object classification.

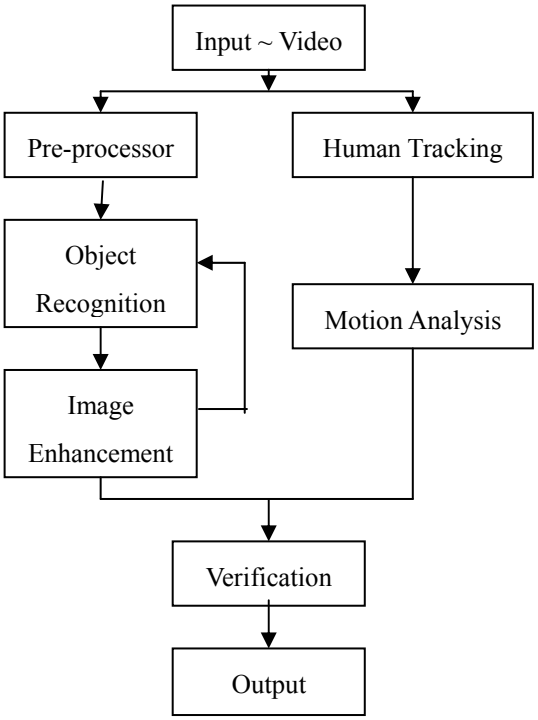


Fig. 2: System Overview

Here, epitome analysis is formulated as the problem of finding the position of the object in V, where V = (μ, ø) and mapping, T_k. In learning an epitome, for each randomly generated training set that contains the object pixels, the value of the parameters, V and T_k of the object are recorded.

So, given the video epitome V = (μ, ø), and the mapping T_k, the watch can be generated by copying the appropriate pixels from the epitome mean and adding Gaussian noise of the level given in the variance map,

$$p(Z_k|T_k,e) = \prod_{i \varepsilon S_k} N(\boldsymbol{Z}_{i,k}; \mu_{T_k(i)}, \varnothing_{T_k(i)}) \quad (9)$$

Although the quality of reconstructed multi-frame images are relatively low, but it still contains the shape property of the object (Figure 5(d)). This shape property data is then compare with the classification result from human motion analysis to avoid the issue of misclassified. That is, for a correct classified system, the shape property achieved from video epitome must be in correlation to the classification result from human motion analysis.

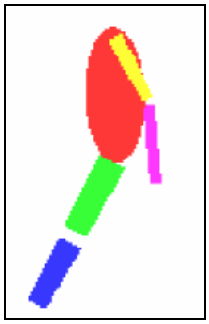


Fig. 3: Example of Kinematic Chain Model of the Human Body used for Tracking

5. Experimental Result

5.1 Experiments Setup and Tracking Initialization

In this paper, the authors captured a single view image size of 120*120 as initial experiments and then extended to a single view video of size 120*160*69. Both image and video are consisted of a man wearing a watch on his arm. The man in the image is static and non static in the video. In the second experiments, the man is detected and tracked using [11] (Figure 6). The result from human body tracking is then passed to the system for a preliminary guess on the object class.

(a) (b)

Fig. 4: (a) Input Image of 120*80 (b) Generated Image Epitome of 50*50 (enlarge by 2 times)

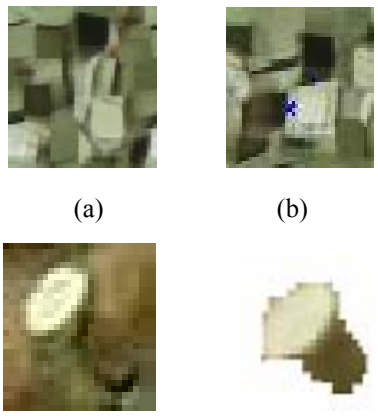
5.2 Image and Video Epitome

In both experiments, the image and video epitome are learned by following procedures. First, the authors will decide on the epitome size where selected epitome size, $V_{x,y,t}$ must be smaller than the input image/video, $I_{x,y,t}$. So in this paper, the authors selected an image epitome with size of 50*50 and 60*45*9 for a video epitome, both epitome size are typically 2.5times smaller that an input image/video. Secondly, randomly generated patch size of 9*9 (for image) and 8*8*2 (for video) are chosen as training data, Then, 10 iterations are selected to learn the epitome. The mean and variance values for the watch in epitome are recorded.

5.3 Test Result

Both experimental results are shown in Figure 5 and Figure 7 where distributed location of the watch in the epitome is represented by blue dots (Figure 5(b) and Figure 7(b)). So, by grouping this information together, a single frame of reconstructed watch is achieved in image epitome (Figure 5(d)) while in video sequence, a multiple frames of reconstructed watch are achieved (Figure 7(c)). Although the reconstructed frame quality is relatively low, but the shape property of the watch is still clearly visible. So, in later stage, the shape property of the watch is passed to the system to compare with object classification result achieved from the aid of human motion analysis. In this experiment, the final result showed that the object is correctly classified in the system.

For the first experiment, the system used 1,800 seconds to perform the task while in second experiment, the system used 2,636 seconds to complete the task. In both cases, convergence is reached in within 5 iterations although 10 iterations were chosen.



(a) (b)

(c) (d)

Fig. 5: (a) Image Epitome of size 50*50 without object recognition; (b) Image Epitome of size 50*50 with object recognition where blue dot indicates watch location distributed in epitome; (c) Original image of a watch (enlarge by 4 times); (d) Reconstructed of the watch using information in (b) (enlarge by 4 times)

6. Further Work and Conclusion

The authors introduced a novel method for human-motion involved object recognition systems in a video sequence. Instead of passing a full length video for analysis, the authors used a method described in [16] where the input video is compressed but still contains the essence of the epitome. Firstly, the size of video epitome is significant smaller than a video. This leads to less computational complexity hence allows applying in application where computational resources are limited. Secondly, it can be trained directly on corrupted or degraded data, as long as the data is repetitive.

Many features of this system are not exploited yet. Besides using an algorithm to find the location of watch in epitomes, experiments are in progress to use a

Fig. 6: Contextual Information from Human Motion Analysis [11]

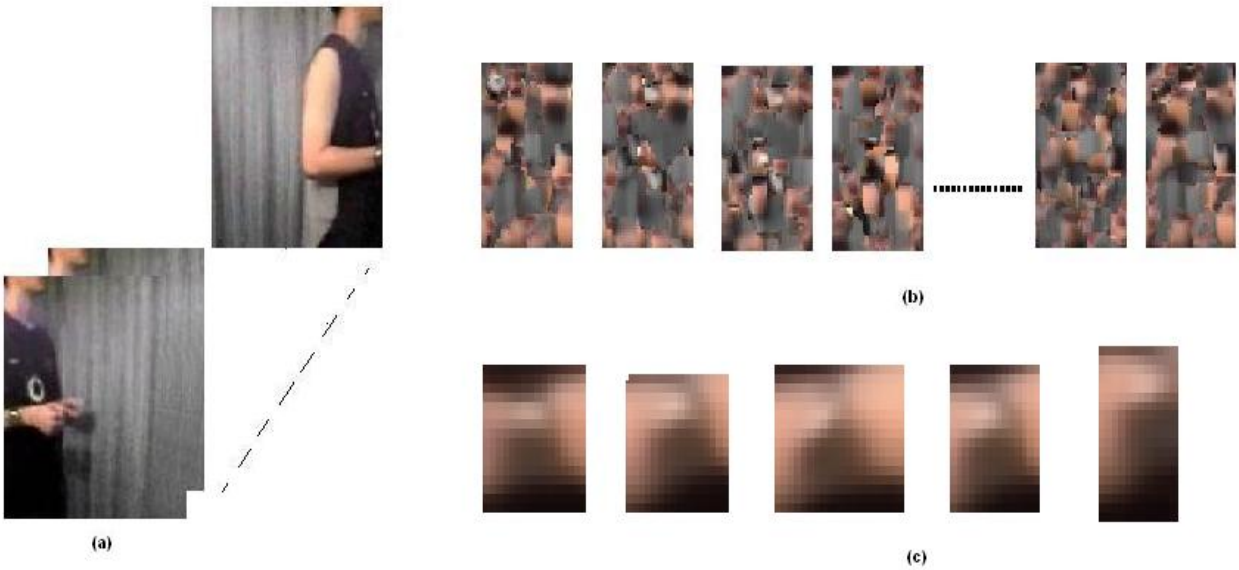


Fig. 7: (a) Input Video of size 120*160*69; (b) Video Epitome size of 45*60*9; (c) Multi-frames Reconstruction (zoom in and enlarge by 4 times)

neural network to solve the matters. Ultimately the system should also be able to account and classify more objects and analysis on different human body part. To apply this technique, a multi-view approach is needed as single view suffers from restriction on viewing angle.

7. Acknowledgement

The authors would like to thank all members of Computer Intelligence & Applications Research Group for helpful discussion and Jiota Research Ltd, UK for incorporating and supporting this research.

References

[1] Isard, I. and MacCormick, J., "BraMBLe: A Bayesian Multiple-Blob Tracker", Proc. of Int. Conf. on Computer Vision, (2001).

[2] Leung, M.K. and Yang, Y.H., "First Sight: A Human Body Outline Labeling System", IEEE Trans. on Pattern Recognition and Machine Intelligence, Volume (17), No.4, (1995), pp.359-377.

[3] Rittscher, J., Kato, J., Joga, S. and Blake, A., "Probabilistic Background Model for Tracking", ECCV, Volume (2), (2000), pp.336-350.

[4] Zhao, T. and Nevatia, R., "Stochastic Human Segmentation from a Static Camera", IEEE Workshop on Motion and Video Computing, (2002).

[5] MacCormick, J. and Isard, I., "Partitioned Sampling, Articulated Objects, and Interface-quality Hand Tracking", ECCV2000, (2000), pp. 3-19.

[6] Lee, M.W. And Cohen, I., "Human Body Tracking with Auxiliary Measurement", IEEE International workshop on Analysis and Modeling of Faces and Gestures, (2003), p.112.

[7] Ju, S.X., Black, M.J. and Yacoob, Y., "Cardboard People: A Parameterized Model of Articulated Motion", (1996), pp. 38-44.

[8] Lipton, A.J., Fujiyoshi, H. and Patil, R.S., "Moving Target Classification and Tracking from

Real-time Video", IEEE Workshop on Application of Computer Vision, (1998).
[9] Goncalves, L., Di Bernadi, E., Ursella, E. and Perona, P., "Monocular Tracking of the human arm in 3D", ICCV, (1995).
[10] Swain, M. and Ballard, D., "Color Indexing", International Journal of Computer Vision, Volume (7), No.1, (1991), pp.11-32.
[11] Bissacco, A., "Visual Tracking of Human Body with Deforming Motion and Shape Average",
[12] Sidenbladh, H., Black, M.J. and Sigal, L., "Implicit Probabilistic Models of Human Motion for Synthesis and Tracking", ECCV2002, Volume (1), (2002), pp. 784- 800.
[13] Sidenbladh, H. and Black, M.J., "Learning Image Statistic for Bayesian Tracking", ICCV, (2001), pp.709-716.
[14] Aggarwal, J.K., and Cai, Q., "Human Motion Analysis: A Review", (1999).
[15] Kakadiaris, I., and Metaxas, D., "Model-based Estimation of 3D Human Motion with Occlusion based on Active Multi-viewpoint Selection", CVPR, (1996), pp.81-87.
[16] Cheung, V., Frey, B.J. and Jojic, N., "Video Epitome", Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (2005).
[17] Fablet, R. and Black, M.J., "Automatic Detection and Tracking of Human Motion with a View-Based Representation", ECCV, Volume (1), (2002).
[18] Fujiyoshi, H., Lipton, A. and Kanade, T., "Realtime Human Motion Analysis by Image Skeletonization", IEEE Workshop on Application of Computer Vision, (1998).
[19] Chen, Z., and Lee, H.J., "Knowledge-guided Visual Perception of 3D Human Gait from a Single Image Sequence", IEEE Trans. Systems, Volume (22), No.2, (1992), pp.336-342.
[20] Deutscher, J. North, B., Bascle, B. and Blake, A., "Tracking through Singularities and Discontinuities by Random Sampling", ICCV, (1999), pp.1144-1149.
[21] Metaxas, D. and Terzopoulos, D., "Shape and Nonrigid Motion Estimation through Physics-Based Synthesis", IEEE Trans. on Pattern Analysis and Machine Intelligence, Volume (15), No.6, (1993), pp.580-591.
[22] Aggarwal, J.K., Cai, Q., Liao, W. and Sabata, B., "Articulated and Elastic Non-Rigid Motion: A Review", Proc. of IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects, (1994), pp.16-22.
[23] Lowe, D.G., "Object Recognition from Local Scale-Invariant Features", ICCV99, (1999).
[24] Haritaoglu, S., Harwood, H. and Davis, L.S., "W^4: Real-Time Surveillance of People and Their Activities",
IEEE Trans. on PAMI, Volume (22), No.8, (2000).
[25] Isard, M. and Blake, A., "Contour Tracking by Stochastic Propagation of Conditional Density", ECCV, (1996), pp.343-356.
[26] Isard, M. and Blake, A., "CONDENSATION - Conditional Density Propagation for Visual Tracking", International Journal of Computer Vision, (1998).
[27] Sullivan, J., Blake, A. and Rittscher, J., "Statistical Foreground Modeling for Object Localization", ECCV, (2000), pp.307-323.
[28] Deutsher, J., Blake, A, and Reid, I., "Articulated Motion Capture by Annealing Particle Filtering", Proc.CVPR, (2000), pp.126-133.
[29] Sidenbladh, H., Black, M. and Fleet, D., "Stochastic Tracking of 3D Human Figures Using 2D Image Motion", Proc. of ECCV, (2000), pp.307-323.
[30] Gavrila, D.M. and Davis, L.S., "Towards 3-D Model-based Tracking and Recognition of Human Movement: A Multi-view Approach", International Workshop on Automatic Face and Gesture Recognition. (1995).
[31] Wren, C., Azarbayejani, A., Darrell, T. and Pentland, A., "Pfinder: Real-Time Tracking of the Human Body", PAMI, Volume (19), No.7, (1997), pp.780-785.
[32] Murase, Hiroshi, and Nayar, S.K., "Visual Learning Recognition of 3D Objects from Appearance", International Journal of Computer Vision, Volume (14), No.1, (1995), pp.5-24.
[33] Shio, A. and Sklansky, J., "Segmentation of People in Motion", Proc. of IEEE Workshop on Visual Motion, (1991), pp.325-332.
[34] Neal, R.M. and Hinton, G.E., "A New View of the EM Algorithm that Justifies Incremental, Sparse and Other Variant", Learning in Graphical Model, (1998).
[35] Choo, K. and Fleet, D.J., "People Tracking with Hybrid Monte Carlo", ICCV2001, Volume (2), (2001), pp.321-328.
[36] Sullivan, J. and Rittscher, J., "Guiding Random Particles by Deterministic Search", ICCV2001, Volume (1), (2001), pp. 323-330.
[37] Jojic, N., Frey, B.J. and Kannan, K., "Epitome analysis of appearance and shape", Proc. of ICCV2003, (2003).
[38] Schiele, B. and Crowley, J.L., "Probabilistic Object Recognition using Multidimensional Receptive Field Histograms", International Conference on Pattern Recognition, (1996).
[39] Marr, D., Vision, W.H.Freeman, 1996.
[40] Schiele, B. and Crowley, J.L, "Object Recognition using Multidimensional Receptive Field Histograms", ECCV1996, Volume (1), (1996), pp.610-619.