Max–Planck–Institut für biologische Kybernetik
Max Planck Institute for Biological Cybernetics

Technical Report No. 174

Combining Appearance and Motion for Human Action Classification in Videos

Paramveer S. Dhillon,[1] Sebastian Nowozin,[2]
and Christoph H. Lampert[2]

August 2008

[1] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, U.S.A, [2] Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany

Combining Appearance and Motion for Human Action Classification in Videos

Paramveer S. Dhillon, Sebastian Nowozin and Christoph H. Lampert

Abstract. We study the question of activity classification in videos and present a novel approach for recognizing human action categories in videos by combining information from appearance and motion of human body parts. Our approach uses a tracking step which involves Particle Filtering and a local non - parametric clustering step. The motion information is provided by the trajectory of the cluster modes of a local set of particles. The statistical information about the particles of that cluster over a number of frames provides the appearance information. Later we use a "Bag of Words" model to build one histogram per video sequence from the set of these robust appearance and motion descriptors. These histograms provide us characteristic information which helps us to discriminate among various human actions and thus classify them correctly.

We tested our approach on the standard KTH and Weizmann human action datasets and the results were comparable to the state of the art. Additionally our approach is able to distinguish between activities that involve the motion of complete body from those in which only certain body parts move. In other words, our method discriminates well between activities with "gross motion" like running, jogging etc. and "local motion" like waving, boxing etc.

1 Introduction

Classification of human actions in video sequences has been extensively studied by the vision community due to its wide number of applications which include video surveillance, object level video summarization, video indexing etc. Besides this, it also provides a summary of the complex video data enabling efficient processing for high level tasks. This task is challenging because of occlusion, background clutter, camera motion etc.

Broadly, the existing approaches can be divided into two categories i.e. supervised and unsupervised. Since our algorithm works in the supervised setting, so we will cover the unsupervised approaches only briefly.

One class of approaches deal with using spatio temporal features for the classification of human actions. In [1] the authors apply spatio temporal volumetric features that scan the video in space and time. Laptev et. al. [2] extend the concept of 2D Harris and Förstner [3, 4] interest point detectors to 3D where the third dimension is temporal. Their approach is based on the assumption that interesting events in videos are characterized by strong variations in both spatial and temporal domains, hence they correspond to spatio temporal interest points "corners". Intuitively a spatio temporal corner is an image region containing a spatial corner whose velocity vector is reversing direction i.e. accelerating. The main assumption of this approach is that accelerating motion is interesting. But, this may not be justified as constant velocity or gradually changing motion may also be interesting [5]. Dollár et. al. propose an alternative approach to behavior recognition using sparse spatio temporal interest points by applying separate filters in spatial and temporal domains (2D Gaussian smoothing kernel and quadrature pair of 1D Gabor filters) respectively. The problem with their approach is that their detector fires only when there is a periodic motion, in other words they are assuming that only periodic events are interesting, which may not be always true and may constrain the performance of the detector when used on a wide variety of video sequences.

Another class of approaches that is commonly used, and which are more closely related to our work, is to perform action classification by tracking a number of spatial features. The authors of [6] use view invariant aspects of the trajectory of a tracked hand to differentiate between various actions. In [7, 8] the authors use the framework of tracking as repeated recognition. Using these approaches the recovery of pose and configurations of the human silhouettes is possible. However, they are still unreliable in domains with cluttered background or in cases where the background has poor contrast.

Lastly, there are approaches which classify human action categories in unsupervised settings. In [9] the authors use spatio temporal words and generative graphical models (pLSA) to learn and recognize human actions in videos.

In this paper we propose to combine appearance and motion information for human action classification in videos. Particle Filters and local non - parametric clustering of particles abstract the appearance and motion information. The approach is based on non Gaussian tracking of the human body parts through a sequence of frames and then extracting the appearance and motion descriptors. The motion information is provided by the trajectory of the cluster modes of a local set of particles. The statistical information about the particles of that cluster over a number of frames provides the appearance information. Later we combine these appearance and motion descriptors to build one histogram per video sequence using a "Bag of Words" model. These highly discriminative histograms provide us characteristic information which helps us to distinguish among various human actions and hence classify them properly. The idea of using tracked motion trajectories of human body parts has been previously used by [10, 11, 12]. But these approaches were highly dependent on the performance of the tracker for their robustness and required much supervision. Our approach differs from these approaches in the fact that we use robust statistical information about the motion of the particles in combination with the information about the trajectory of the cluster modes for the classification of human actions. As the approach does not completely rely on one aspect of tracking and since the histogram representation is highly robust to clutter hence performance of the tracking subroutine ceases to be a bottleneck in the performance of our approach.

The rest of the paper is organized as follows. In Section 2 we briefly review Particle Filters, then we describe our approach in detail, including the various statistical models that our system uses in Section 3. In Section 4 we present the experimental results of our algorithm on the KTH [13] and Weizmann [14] human action datasets. Finally we conclude in Section 5 by giving a summary of our approach.

2 Mathematical Background

We use Sequential Monte Carlo methods/Particle filters for tracking which considers x_t as the state variable (like position, velocity etc.) and y_t as the actual measurements/observations made in the image.

2.1 Bayesian Filtering and Sequential Monte Carlo Methods Revisited

Given an internal state sequence $\{\mathbf{x_t}; \mathbf{t} \in \mathbb{N}\}, \mathbf{x_t} \in \mathbb{R}^{\mathbf{n_x}}$ and an observation sequence $\{\mathbf{y_t}; \mathbf{t} \in \mathbb{N}\}, \mathbf{y_t} \in \mathbb{R}^{\mathbf{n_y}}$ where n_x is the dimension of the internal state vector and n_y is the dimension of the observation vector, the Bayesian Filtering distribution can be written in two steps as follows:

- **Prediction:**
  $p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1}$
- **Update:**
  $p(x_t|y_{1:t}) = \frac{p(y_t|x_t)p(x_t|y_{t-1})}{\int p(y_t|x_t)p(x_t|y_{1:t-1})dx_t}$

where
$p(x_t|x_{t-1})$ is the transition distribution
$p(x_t|y_{1:t})$ is the posterior distribution at current time step

SMC (Sequential Monte Carlo Methods) or Particle Filters provide an approximate solution to the above two recursive equations by using a large set of random samples called particles. In standard particle filters we approximate the posterior $p(x_t|y_{1:t})$ with a set of Dirac functions centered at finite set of N particles $\left\{\mathbf{x_t^i}\right\}_{i=1...N}$:

$$p(x_t|y_{1:t}) \approx \sum_{i=1}^{N} w_t^i \delta_{x_t^i}(\mathbf{x_t}), \tag{1}$$

where w_t^i is the weight associated with the i^{th} particle and is calculated as:

$$w_t^i \propto w_{t-1}^i \frac{p(y_t|x_t^i)p(x_t^i|x_{t-1}^i)}{q(x_t^i|x_{t-1}^i, y_t)}, \tag{2}$$

where q(.) is the proposal distribution or the importance density function which is often chosen to be $p(x_t|x_{t-1})$. Very often a resampling algorithm is applied to avoid the degeneracy problem [15], in which case $w_{t-1}^i = \frac{1}{N} \forall i$, hence

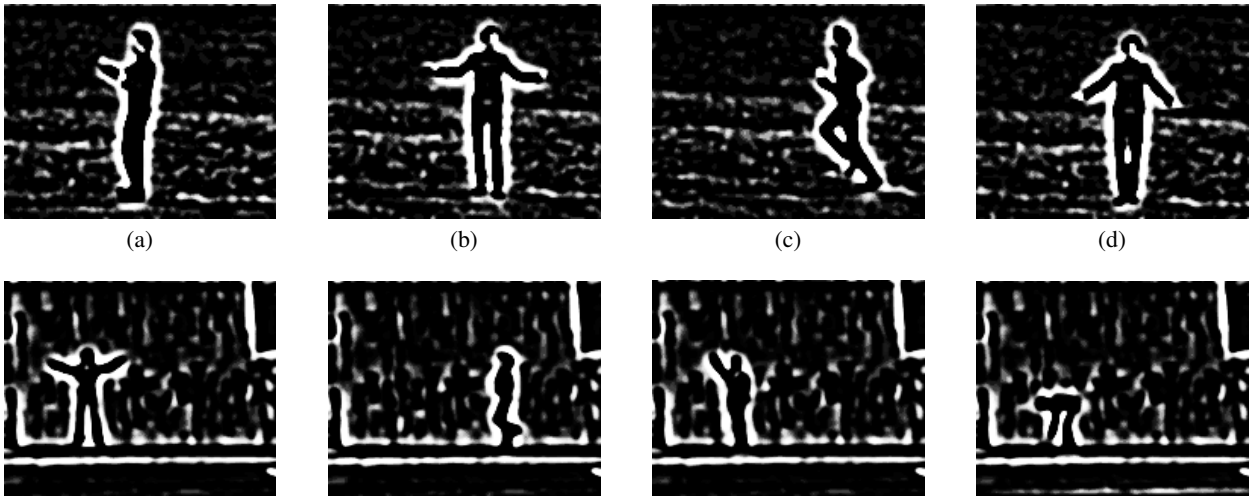


Figure 1: Areas of Strong Filter Response: (a) - (d) shows activities [boxing, handclapping, running, handwaving] from the KTH dataset [13] and (e) - (h) shows activities [jack, jump, wave, bend] from the Weizmann dataset [14]

$$w_t^i \propto p(y_t|x_t^i), \tag{3}$$

i.e. the weights are proportional to the likelihood function. The resampling step derives the particles depending on the likelihood function of the previous step, and all the particles receive a starting weight equal to $\frac{1}{N}$ which will be updated by the next frame likelihood function. The major advantage of using particle filters is that they can handle multimodal likelihoods [16] and can track even after an occlusion.

3 Our Approach

3.1 Statistical Model

In order for a particle filter to work we need to chose an initial distribution $p(x_0)$ and a transition distribution $p(x_t|x_{t-1})$ for our state sequence $\{x_t; t \in \mathbb{N}\}$. Besides this we also need to chose an observation / likelihood model which specifies the likelihood of an object being in a specific state. The observations $\{y_t; t \in \mathbb{N}\}$ are conditionally independent given the state $\{ x_t; t \in \mathbb{N}\}$, with marginal distribution $p(y_t|x_t)$. We now describe the models we use.

3.1.1 Initial Distribution $p(x_0)$ and Transition Distribution $p(x_t|x_{t-1})$

In our case the initial distribution for the spread of particles consists of the areas of high response of the spatial interest point operator [17]. We squash the output of a DoG (Difference of Gaussians) filter by a sigmoidal. These spatially strong response regions are the regions which are most likely to capture the "interestingness" of the frame [2, 5] and hence will serve as a good prior for initial distribution of particles. The areas of strong filter response for some activities in the KTH [13] and Weizmann [14] datasets are shown in Fig. 1.

As far as the transition distribution $p(x_t|x_{t-1})$ is concerned, we use a standard second order autoregressive model [18] as below:

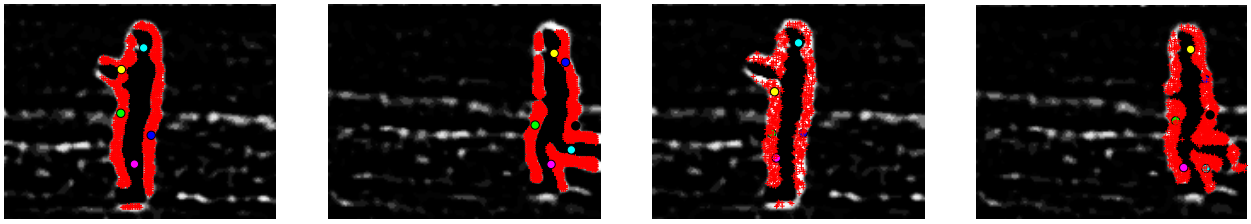
$$X_t - X_0 = A(X_{t-2} - X_0) + B(X_{t-1} - X_0) + C\epsilon_t \tag{4}$$

where A,B and C are constants and ϵ_t is Gaussian random noise used to diffuse the particles.

Robustness of this model comes from the fact that it takes into account previous states for velocity and acceleration information and hence captures the dynamics of the motion.

3.1.2 Observation Model $p(y_t|x_t)$

The observation model is one of the most important factors which determines the performance of the tracker. There have been a variety of observation models that have been used in literature depending on the application and data. The most common ones being multi-color observation models based on Hue Saturation and Value (HSV) color histograms [19]. They are widely used as they are reasonably insensitive to illumination effects as HSV



(a) (a′) (b) (b′)

Figure 2: Plots for steps 3 - 5: (a - a′) Shows the initial distribution of particles and the cluster modes [boxing and running (KTH dataset)] , (b - b′) Shows the drifted and diffused particles and cluster modes

decouples the intensity (i.e. value) from color (i.e. hue and saturation). Models based on boosting are also commonly used [18].

We investigated likelihood models for gray-scale images, as such models are simple compared to their color counterparts and can speed up the computation, without sacrificing much of the performance.

- **A Global Likelihood Model:**
  This model is based on the assumption that the predicted position of the particle in the next frame should also lie in spatially interesting region. We measure the likelihood considering a n $\times$ n window around the particle position (y_t).

  $$p(y_t|x_t) \propto \textit{Number of ``interesting'' pixels in that n} \times \textit{n region surrounding the particle} \tag{5}$$

  Due to its simplicity, this model is computationally efficient and at the same time does not sacrifice the performance.

- **A Local (Gaussian) Model:**
  This model postulates that the predicted position of the particle in the next frame should be similar in appearance to its position in the current frame. It can be represented as:

  $$p(y_t|x_t) \propto e^{-\lambda \sum_{i=1}^{n^2} |I_t^i - I_{t-1}^i|^2} \tag{6}$$

  where I_t^i are the normalized pixel values in frame t for the particle i.

3.2 Overview of the method

Our algorithm is summarized in Algorithm 1. We extract a characteristic video histogram for each video sequence based on tracking information obtained from the motion of the particles and tracked motion trajectories of human body parts. The basic idea is that human activities can often be characterized rather well by appearance and the local motion of the body parts.

At the beginning we spread the particles uniformly on the high response areas of a DoG (Difference of Gaussian) filter. Intuitively this corresponds to a region which has "spatial interestingness" and hence serves as a good prior for the initial distribution of the particles. As a next step, the particles are clustered by Mean Shift Clustering [20] (since we do not know the number of clusters *a priori*) and only sufficiently dense clusters (containing ≥ 2 % of total particles) are kept as can be seen in Fig. 2. The cluster modes correspond to the regions of high particle density. The clustering step allows us to associate a fraction of particles to each cluster mode. In essence we have a mixture particle filter [21] i.e. a separate particle filter for each cluster mode. Besides this it also allows us to vary the granularity of the information to be extracted by choosing a low or high value of the kernel bandwidth. The cluster modes provide rich information about the relative motion of the particles belonging to that cluster, so we chose to attach a local log-polar histogram to each cluster mode to extract this appearance information quantitatively.

In our approach, the log-polar histogram (5 radial and 12 angular bins) attached to each cluster mode, provides information about appearance. The motion information is provided by the *trajectory* of the cluster mode. We get a set of robust descriptors from the appearance and motion information that can not only discriminate among various activities (based on appearance) but can also distinguish between slow and fast activities (based on motion of the cluster mode).

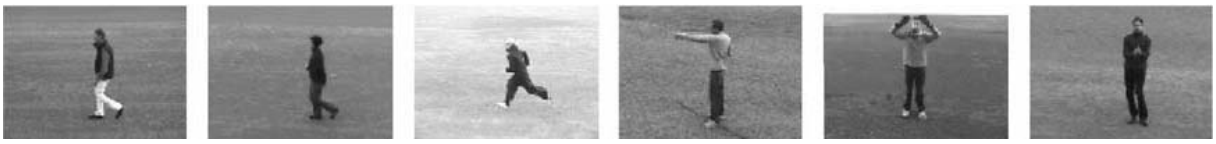


Figure 3: Example sequences from KTH dataset [walking, jogging, running, boxing, handwaving and handclapping]

We combine these set of appearance and motion descriptors into one histogram by a "Bag of Words" representation. We cluster the descriptors by k-means to obtain a codebook and then quantize the descriptors into the bins of the codebook vectors for each video sequence. Finally these histograms are classified by a SVM [22]. For simplicity and speed we use a linear SVM.

As is obvious, in our model we do not make any implicit assumptions about the interestedness of the motion as is done by [2, 5].

Algorithm 1 Human Action Classification in Videos

1: **for** {videos= 1: endVideo} **do**
2: **for** {frames=1: endFrame} **do**
3: Distribute the particles on the squashed response of the spatial interest point detector i.e. the DoG filter.
4: Cluster the particles locally by using Mean Shift Clustering. Attach a log-polar binned histogram to each cluster mode.
5: Drift, diffuse and resample the particles. Update the cluster modes based on the particles belonging to that cluster.
6: Obtain mean of the motion of the particles in the bins of the histograms. Also obtain the trajectories of the cluster modes.
7: **end for**
8: Use "Bag of Words" representation to build appearance and motion histograms.
9: Normalize and combine these histograms to get one histogram per video and classify it using a linear SVM classifier.
10: **end for**

The plots for the Steps 3, 4 and 5 of Algorithm 1 are shown in Fig. 2.

4 Experimental Results

In this section we present results on two datasets: KTH human motion dataset [13] and Weizmann human action dataset [14]. Each dataset contains videos of cluttered background, moving cameras and multiple actions. The datasets and the results are explained in detail in the following sub-sections.

4.1 KTH Human Motion Dataset

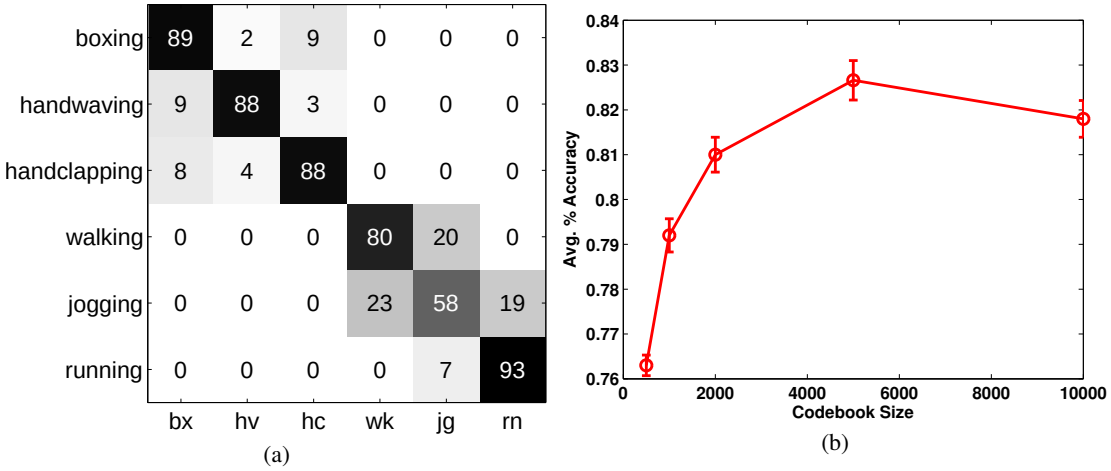
KTH Dataset is the largest available and most standard dataset used for benchmarking results for human action classification. The dataset contains six activities (boxing, handwaving, handclapping, running, jogging, walking) performed by 25 subjects in 4 different conditions (like illumination, background, zoom). The dataset contains a total of 598 video sequences and in each video only one action is performed. Some sample sequences are shown in Fig. 3.

We build one normalized histogram per video by combining our set of motion and trajectory descriptors. For classification purposes we use a linear SVM. We perform 10 fold CV on the training set to fine tune our parameters and then perform training and testing on the sequences as mentioned on the dataset homepage [13]. As the approach is probabilistic we run the experiments 10 times and report the mean and the standard deviation.

For the "Bag of Words" model we experiment with different codebook sizes varying from 500 to 5000. The variation of performance as a function of codebook size is plotted in Fig. 4(b). It was observed that in most cases the classifier became much more discriminative with increase in size of the codebook.

To evaluate the performance of our approach we plot the results in the matrix form showing the confusions among various activities. The confusion matrix for the six classes of the KTH dataset is shown in Fig. 4(a).

As is obvious from the results there is confusion between 'walking' and 'jogging' and also between 'boxing' and 'handclapping', which is quite intuitive as these actions are quite similar and even seeing the video sequences with naked eye, we cannot fully distinguish between these activities. Also, another thing worth mentioning is that, our approach is able to distinguish between activities with local and gross motion. This fact is obvious from



(a) (b)

Figure 4: Results for the KTH dataset (using the Local likelihood model) (a) shows results for a linear SVM classifier, (b) shows the average accuracy as a function of codebook size.

Figure 5: Example sequences from Weizmann dataset [14] [bend, p-jump, wave2, run, jump, jack, walk, wave1, skip and side]

our results, as they bring out a better distinction between gross motion (like jogging, running, walking) and local motion (like boxing, handwaving, handclapping).

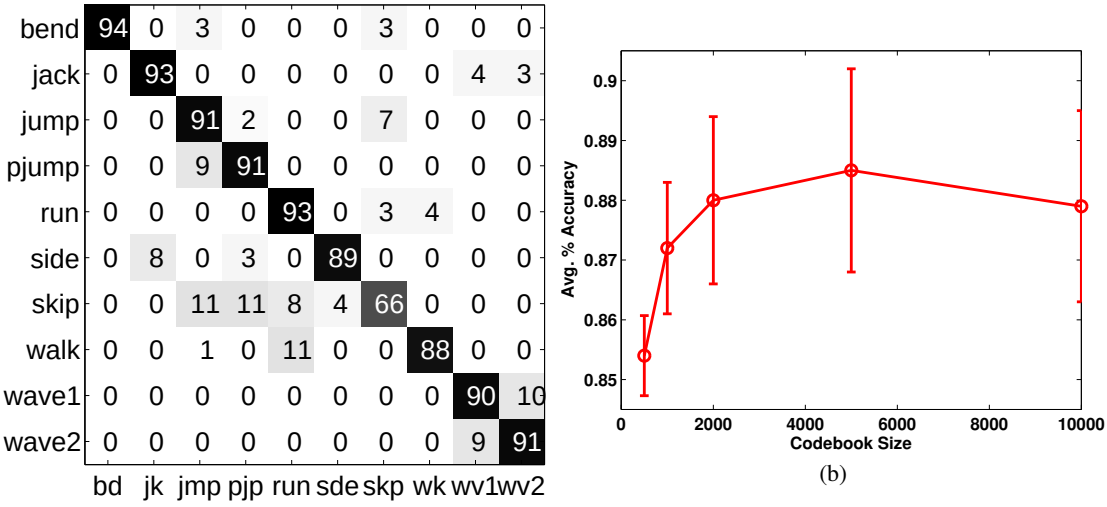
On average the performance was 82.66 %, for all activities, with a standard deviation of 0.44 % for the linear SVM classifier.

4.2 Weizmann Human Action Dataset

The second dataset that we use to test our approach is the Weizmann Human Action Dataset. It contains a total of 10 actions performed by 9 people, to provide a total of 90 videos. Sample sequences are shown in Fig. 5. The dataset contains videos with a static camera unlike KTH, where some of the videos had zooming and also the videos have simple background. However, this dataset contains 10 activities, which is more compared to 6 activities of KTH dataset, so it will provide a good test to our approach in the setting in which the number of activities are increased.

Again, we do 10 fold CV on the training set to fine tune our parameters and then make a 70 - 30 split for training and testing set. The classifier used is the same as earlier i.e. a linear SVM. The confusion matrix is shown in Fig. 6(a) and the performance dependence on codebook size is show in Fig. 6(b).

As is obvious from the confusion matrix, there was some confusion in the classification of the “skip” activity. Overall the mean accuracy was 88.5 % averaged over 10 runs, with a standard deviation of 1.4 %.



(a) (b)

Figure 6: Results for the Weizmann dataset (using the Local likelihood model) (a) shows results for a linear SVM classifier, (b) shows the average accuracy as a function of codebook size.

5 Summary and Conclusion

In this paper we proposed a novel approach for human action classification in videos based on combining appearance and motion information obtained from tracking of the human body parts. A mixture particle filter [21, 18] approach is used in which a single filter is responsible for each cluster mode. The approach is based on extracting a set of descriptors based on appearance and motion. We get the appearance information from the motion of the particles in the bins of the log-polar histogram attached to each cluster mode and the motion information from the trajectory of the individual cluster modes. It turns out that the histograms obtained from the bag of words representation of these descriptors are not only characteristic of the activity being performed but are also highly robust to clutter, as can be seen by the performance on Weizmann dataset in which there was lot of background clutter.

We demonstrated results on two standard datasets i.e. the KTH and Weizmann Human Action Datasets. The results were comparable to the state of the art and they validate our proposed approach. Besides this, we also proposed two likelihood/ observation models for gray-scale images which are simple and easy to implement.

In future we would like to experiment with other likelihood models for gray - scale images and would like to extend our approach to detect multiple activities in the same video and test it on a more challenging dataset.

References

[1] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 166–173, Washington, DC, USA, 2005. IEEE Computer Society.

[2] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 432, Washington, DC, USA, 2003. IEEE Computer Society.

[3] C. Harris and M. Stephens. A combined corner and edge detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[4] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. *In Intercommission Conf. on Fast Processes of Photogrammatic Data*, pages 281–305, 1987.

[5] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

[6] C. Rao and M. Shah. View-invariance in action recognition. In *CVPR01*, pages II:316–322, 2001.

[7] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *ECCV02*, page I: 629 ff., 2002.

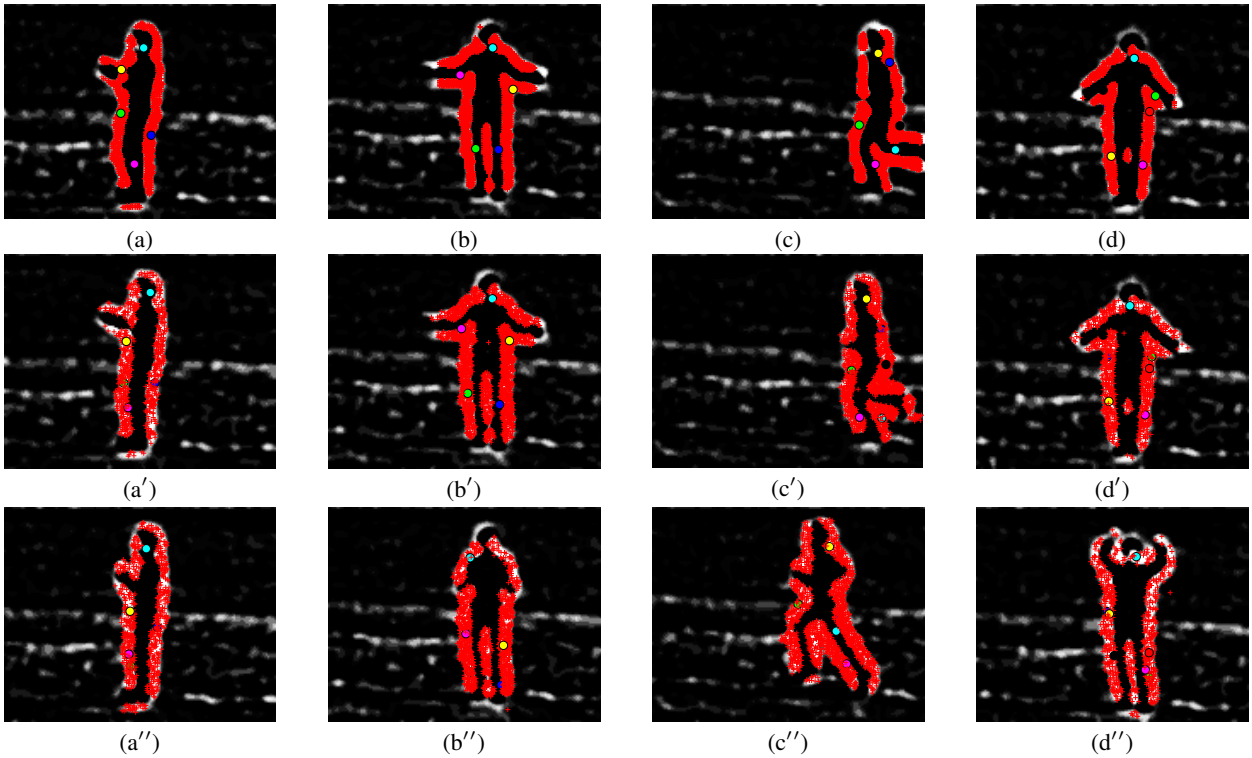


Figure 7: More Results from KTH Dataset: Figures (a) - (d) show various activities with particles distributed and the initial cluster mode locations and (a′) - (d′) and (a′′) - (d′′) show the drifted, displaced particles and cluster modes in later frames. This figure is best viewed with color. Note: Only sufficiently clusters are kept, rest are culled.

[8] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *ECCV (3)*, pages 666–680, 2002.

[9] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 2008.

[10] A. Agarwal and B. Triggs. Learning to track 3d human motion from silhouettes. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 2, New York, NY, USA, 2004. ACM.

[11] D. Ramanan and D. A. Forsyth. Automatic annotation of everyday movements. In *NIPS, 2003*, 2003.

[12] A. Yilmaz and M. Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 150–157, Washington, DC, USA, 2005. IEEE Computer Society.

[13] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3*, pages 32–36, Washington, DC, USA, 2004. IEEE Computer Society.

[14] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 1395–1402, Washington, DC, USA, 2005. IEEE Computer Society.

[15] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[16] M. Isard and A. Blake. Condensation — conditional density propagation for visual tracking. *IJCV*, 29:5–28, 1998.

[17] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[18] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV (1)*, pages 28–39, 2004.

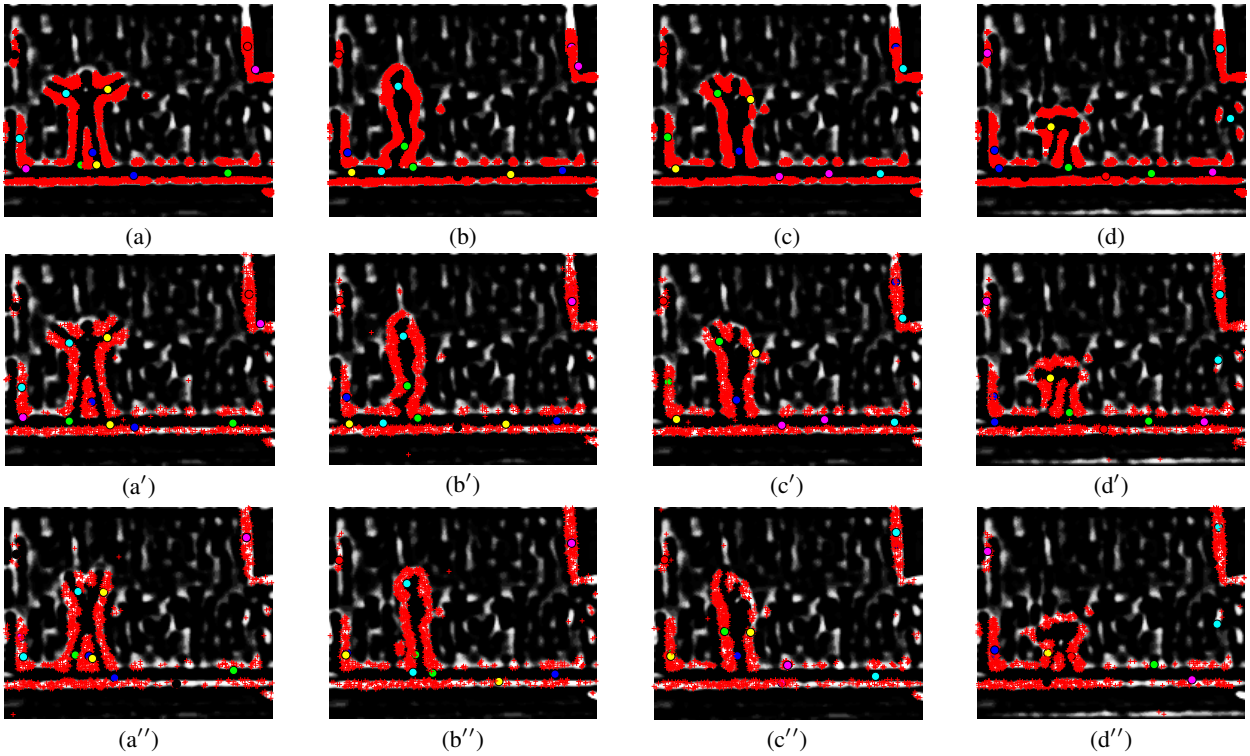


Figure 8: More Results from Weizmann Dataset: Figures (a) - (d) show various activities with particles distributed and the initial cluster mode locations and (a′) - (d′) and (a′′) - (d′′) show the drifted, displaced particles and cluster modes in later frames. As is obvious lots of particles are distributed in background areas but this does not affect the performance of our method as our histogram representation is robust to clutter. This figure is best viewed with color. Note: Only sufficiently dense clusters are kept, rest are culled.

[19] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *ECCV (1)*, pages 661–675, 2002.

[20] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002.

[21] J. Vermaak, A. Doucet, and P. Pérez. Maintaining multi-modality through mixture tracking. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1110, Washington, DC, USA, 2003. IEEE Computer Society.

[22] B. Schölkopf and A. J. Smola. *Learning With Kernels*. MIT Press, Cambridge, MA, USA, 2002.