Action Recognition Using a Spatio-Temporal Model in Dynamic Scenes

K. G. Manosha Chathuramali, Ranga Rodrigo
Department of Electronic and Telecommunication Engineering
University of Moratuwa
Sri Lanka
Email: manosha@ent.mrt.ac.lk, ranga@uom.lk

***Abstract*—Action recognition in a video plays an important role in computer vision and finds many applications in areas such as surveillance, sports, and elderly monitoring. Existing methods mostly rely on stationary backgrounds. Action recognition in dynamic backgrounds typically requires standard preprocessing steps such as motion compensation, background modeling, moving object detection and object recognition. The errors of the motion compensation step and background modelling increase the mis-detections. Therefore action recognition in dynamic background is challenging. In this paper, we use a combination of pose characterized by a silhouette and optic flows synthesized into a histogram. This enables us to classify the movement of the actor versus movement of the background. We use four background models to extract the silhouette from the frame. We use SVM to recognize actions, according to several evaluation protocols. We perform several experiments and compare over a diverse set of challenging videos, including the new Change Detection Challenge Dataset. Our results perform better than existing methods.**

Keywords*—*Dynamic backgrounds, background modeling, AMM, FDM, GMM, JBFM, SVM

I. Introduction

Action recognition in videos is an active research topic in computer vision with many important applications in areas such as surveillance, human-computer interaction, video retrieval, robot learning, sports and elderly monitoring. The capability of automatically recognizing actions in security-sensitive areas such as airports, borders, and building lobbies is of great interest.

There are many approaches that recognize actions in videos with stationary backgrounds. The assumption of background being stationarity, limits the application of these algorithms. However, accurate and reliable action detection from cluttered and highly dynamic background remains a challenging problem. Methods that are robust and generic enough to handle the complexities of the most natural scenes are still very limited [1], [2]. Existing work mainly focuses on evaluating indoor and outdoor surveillance videos which often have relatively stationary backgrounds. However, videos captured in natural environments contain non-stationary background such as swaying tress and rippling water, that have not been sufficiently addressed in the literature [2]. The key challenge here is to incorporate effective models to capture the highly dynamic backgrounds.

There are two main traditional approaches to detect moving objects in dynamic scenes: (1) using motion compensation step by performing video alignment [3], [4], (2) background modeling followed by the tracking of the detected moving blobs [5]–[7]. These approaches face many problems: First, video alignment is difficult and noisy due to the perspective distortions, the errors in feature point detection and localization, and the errors from alignment. Second, the errors of alignment and moving object detection further propagate to the tracking stage. Monnet *et al.* [8] propose an on-line auto-regressive model to capture and predict the behavior of dynamic scenes. In order to detection of events, they introduce a new metric that is based on a state-driven comparison between the prediction and the actual frame. Zhong *et al.* [9] use a foreground-background method along with an auto-regressive average model to characterize dynamic scenes and the Kalman filter to estimates the intrinsic appearance of the dynamic texture and regions of the foreground objects. Sheikh and Shah [10] propose a method called temporal persistence in conjunction with background difference in dynamic scenes. They use Bayesian models for both the background and the foreground to make decisions based on spatial context. These systems have mainly focused to detect moving objects from the highly dynamic backgrounds. If there are two or three objects moving in the same background, it is difficult to recognize different behaviours or appearances of them. However, our intention is to detect actions such as people walking instead of detecting moving objects from the dynamic backgrounds. We use different protocols (test cases) to identify the differences between the moving objects in the same background. Instead of classifying object vs. non-object, we train a multi-class SVM classifier with a rich feature vector to recognize actions, which is similar to detections in stationary backgrounds [11].

In this work, we use a combination of three methods: modeling the background to get the foreground object, creating the local descriptors and combining them to create the motion descriptors. Background modeling is the initial step in the video analysis with dynamic background. We considered both simple models such as frame differencing and approximate mean models, and relatively complex models such as Gaussian mixture model and Bayesian models. This enables us to extract silhouette. It is necessary to consider both motion and pose to extract features to detect actions. This addresses a basic problem in action recognition: what are good features that describe actions. Following the work of Tran *et al.* [11], we use a frame sequence descriptor which is a histogram of optic flow values and the silhouettes. An action spans over a number of frames. In our method, we select 15 frames at a time and project the feature descriptors to form a block descriptor using PCA to a low-dimensional space. Our method is a fair representation of both pose and motion. Finally, we use SVM to classify the feature vectors. In this paper, we present several experiments and performance comparisons over a diverse set of challenging videos, including the recent Change Challenge Dataset [12].

This paper is organized as follows: section 2 gives an overview of our method, in which we describe background models and the feature extraction method. Activity detection and SVM classifier are briefly described in section 3. Descriptions about the dataset that we used and evaluation methodology are given in section 4. Experimental results are shown in section 5 and we compare our results with other methods. Finally, we briefly discuss our approach and future directions in section 6.

II. Methodology

In this paper, we address the problems in detecting actions in dynamic backgrounds using spatio-temporal descriptors. We handle the dynamic background by incorporating a background subtraction in the action detection pipeline. We use four background models. They are Approximate Mean Model (AMM), Gaussian Mixture Model (GMM), adjacent Frame Difference Model (FDM), and Joint Background-Foreground Model (Bayesian modeling approach) (JBFM) as described in Sheikh and Shah [10]. After extracting silhouettes, we extract features from the silhouettes and combine these features with optic flow values and create a rich feature vector [11].

Our feature extraction method comprises three steps: first, using background subtraction get the silhouette of the actor; second, extracting local features from each frame; third, finding global features through an action sequence comprising several frames.

A. Background Subtraction

We use a diverse set of dynamic scenes for our experiments. We use sample frames from few datasets to show the dynamic background in figure 1 using optic flows. These scenes do not include periodic motions. We use four different background models to see the differences in their performance in action detection along with the spatio-temporal features that we use in this work. Although some of these methods are known to give poor results for object detection from the dynamic background, we use them to validate our feature extraction and classification methods. As mentioned before, we use AMM, GMM, FDM, and JBFM background models in this work.

The Bayesian model approach, as described in [10], is a joint background and foreground model, to detect objects in dynamic scenes. Background data is modeled as a single distribution. The foreground is explicitly modeled to augment the detection of objects without using tracking information. Finally, it uses a Bayesian model, that competitively use both the background and the foreground to make decisions based on the spatial context of pixel neighborhood labels.

B. Local Features

A local feature is a histogram of the silhouette of the actor-object and the computed optic flow values. We use frames scaled to 360×240 to extract the silhouette using the aforementioned background subtraction method and to resample the flow vectors. The optical flow vector field F is first split into two scalar fields corresponding to the horizontal and vertical components of the flow, Fx and Fy respectively. To compute the optic flow values we use the Lucas-Kanade algorithm [13]. These three channels, silhouette, F_x, and F_y, are smoothed by a median filter to reduce the effect of noise. Each of these channels is histogrammed using the following technique: we divide the 360×240-scaled frame into 2×2 sub-windows and then each sub-window is divided into an 18-bin radial histogram (20 degrees per bin). The radial histograms perform well in 2×2 sub-windows. However, they are meaningless when the sub-windows are too small. These pie slices do not overlap and the center of the pie is in the center of the sub-window. The values of each channel are integrated over the domain of the slice. The result is a 72-($2 \times 2 \times 18$)-dimensional histogram. By concatenating the histograms of all three channels we get a 216-dimensional frame descriptor [11].

C. Motion Descriptor

Silhouette and optic flow based descriptors described above captive local information. However, action recognition needs to consider temporal information as well. Following Tran *et al.* [11], we too use 15 frames and split them into 3 blocks of 5 frames, named as past, current and future. The frame descriptors of each block are stacked together into a 1080-(72×15)-dimensional vector. Due to the high dimensionality of the feature vector we use principal component analysis (PCA) to reduce the number of dimensions. Then the resulting 70-dimensional context descriptor is appended to the current frame descriptor to form the final 286-dimensional motion context descriptor.

III. Action Recognition

Our objective is to detect actions in the dynamic backgrounds. For example, we use different protocols to detect the difference between “single person walking” and the “two people walking” in the dynamic backgrounds. We use SVM to recognize the actions by using the feature descriptors described in section 2.

We use both multi-class and one-class SVM classifiers to detect actions in the experimental datasets. We use LIBSVM [14] in our experiments. To find the best classifier for our datasets, we carry out a grid search in the space of parameter C and γ. Here, C is the weight of error penalty and γ determines the width of the RBF kernel. The appropriate SVM classifier is selected by the set of (C,γ) which maximizes the cross-validation rate in the space of search, which, in turn, increases the accuracy of the results [15].

IV. Datasets and Evaluation Methodology

A. Datasets

We use diverse sets of existing benchmark datasets that include dynamic backgrounds for our experiments: (a) *WavingTrees* [16]; (b) three datasets (*Fountain*, *Ducks* and *Railway*) [10]; (c) *Boat-sea* [17]; (d) three datasets (*Curtain*, *Watersurface* and *Fountain3*) [18] and (e) five datasets (*Boat*, *Canoe*, *Fountain1*, *Fountain2* and *Overpass*) in new Change

TABLE I: Action recognition rate against the four background model: AMM, FDM, GMM and JBFM. "*" refers to the disregarded datasets when calculating the average.

Dataset	AMM	FDM	GMM	JBFM
Boat	66.81	71.95	61.24	**91.85**
Canoe	51.69	96.19	**100**	**100**
Curtain	81.9	**98.9**	86.09	96.69
Overpass	64.85	45.17	58.10	**69.76**
Watersurface	84.56	91.27	77.18	**100**
Fountain1	67.22	70.54	**89.21**	72.62
Fountain2	87.17	87.18	95.86	**98.46**
Fountain3 (F3-1)	10.34	10.34	13.79	**72.41**
Fountain3 (F3-2)	74.14	60.34	87.93	**100**
*Switchlight (SL-1)	**70.39**	36.84	44.74	54.61
*Switchlight (SL-2)	**54.32**	40.74	28.39	12.34
Average Change Detection Challenge Dataset	67.55	74.21	80.88	**86.54**
*Average	68.31	74.57	77.31	**89.45**

TABLE II: Action recognition rates with dynamic backgrounds in different datasets. See IV B for a description of the protocols.

Dataset	Protocol	Test Seq.	Recog.	Mis-Recog.	Rate
Fountain	F1	51	51	0	100
	F2	78	78	0	100
	F3	21	21	0	100
Ducks	D1	69	69	0	100
	D2	200	194	6	97
Railway	R1	150	150	0	100
	R2	100	100	0	100
	R3	40	39	1	97.5
Boat-sea		175	163	12	93.1
Wavingtress		8	8	0	100

Detection Challenge dataset [12] for performance evaluation and comparisons with existing methods. Selected datasets include different types of dynamic backgrounds of both indoor and outdoor scenes. Figure 2 shows sample frames of these datasets with objects/non-objects.

B. *Evaluation Methodology*

To perform detecting actions in dynamic backgrounds, we use four background models to extract silhouettes and then we create the motion descriptors for each model. We compare the results of these four models. We use first 100 frames to learn the background in JBFM for each dataset. It is a challenging task to experiment with four background models followed by creating motion descriptors for the selected nine datasets, due to the long computational time in creating motion descriptors.

To recognize actions we use different protocols for datasets as described follows. We use two protocols for *Fountain3* dataset: two people walking (F3-1) and one person walking (F3-2) and with the *Switchlight* dataset we use two protocols: one person walking (SL-1) and two people talking (SL-2). With the *Fountain* dataset we use three protocols: single person walking (F1), single person walking in the opposite direction (F2), two or three people walking at the same time (F3). With the *Ducks* dataset we use two protocols: detecting a single duck (D1) and detecting two ducks (D2). With the *Railway* dataset we consider three protocols: a person moving across the railway (R1), a car moving across the railway (R2), and both the car and the person moving across the railway (R3). With the *Boat-sea* dataset we have only one protocol, the movement of the boat (B1). .

V. RESULTS AND DISCUSSION

A. *Quantitative Results*

Table 1 shows the action recognition rate for dynamic background videos in the new Change Detection Challenge dataset [12] and other commonly used datasets against four background models. The average (89.45%) is higher in JBFM. Other two simple models, frame differencing and approximate mean, have good average although they are known to give poor results in the literature. This improvement is probably due to the good motion descriptor that we used. We use the same number of frames for creating feature descriptor, training and testing. This facilitates the comparison of suitability of each background model to be used in an action recognition system. JBFM seems to be better on average. Although average is higher in JBFM, other methods outperform JBFM in some datasets. For example, in *Curtain* dataset best value is 98.9% in FDM, in *Fountain1* dataset best value is 89.21% in GMM. In JBFM, we use the first 100 frames for learning the background and then extract the silhouette. The major drawback here is that variations after 100 frames in the background are detected and learned as foreground. This may be the reason for poor results in some datasets. The other three models have their own drawbacks: FDM cannot detect still people in dynamic backgrounds. Using FDM, it is difficult to detect actions, like the differences between moving objects. It gives poor results for two protocols in *Fountain3* dataset.

We disregarded the *Switchlight* dataset in reporting the average in Table 1 as it did not contain any switching off of lights in the first 100 frames. Therefore we were unable to extract the silhouette using JBFM properly.

Change Detection Challenge dataset [12] website publishes results for an extensive list of methods recently reported in the literature. The rate that we have achieved from JBFM, 86.54%, is higher than the second best 83.26% [19]. This is because both background model and the feature extraction method in our system are able to capture the actions in highly dynamic backgrounds effectively.

In Table 2, we use only JBFM as it is proven to be giving better performance as seen in our previous experiments. We achieve 100% for most of the datasets under various protocols. Our results are superior in detecting actions when compared to the results in [10].

B. *Qualitative Results*

In this section, we provide qualitative results of four background models representative of the different quantitative results achieved in Table 1 and Table 2.

Figure 3 shows the results of the three datasets against the four background models: AMM, FDM, GMM and JBFM. These results clearly show the JBFM extract the silhouette and perform well than other three models. This resulted in better performance in action recognition in JBFM shown in Table 1.

Figure 4 gives a picture of the *Switchlight* dataset and poor results of JBFM model for light off situation as explained in better.

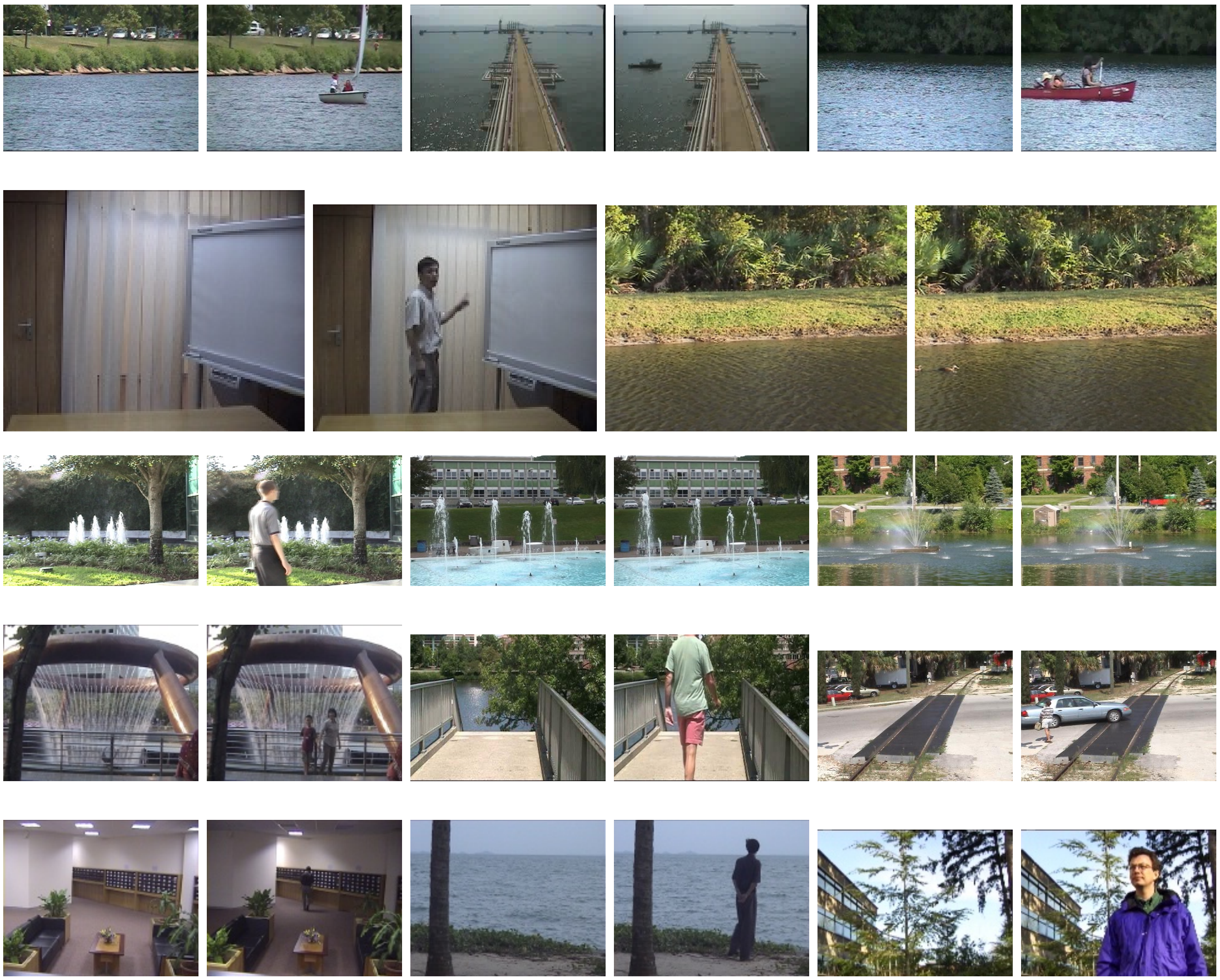


Fig. 2: Sample frames of datasets (dynamic background/dynamic background with object) that we have used for our performance evaluation. The datasets are: *Boat*, *Boat-sea*, *Canoe*, *Curtain*, *Ducks*, *Fountain*, *Fountain1*, *Fountain2*, *Fountain3*, *Overpass*, *Railway*, *Switchlight*, *Watersurface*, *Wavingtrees* respectively.

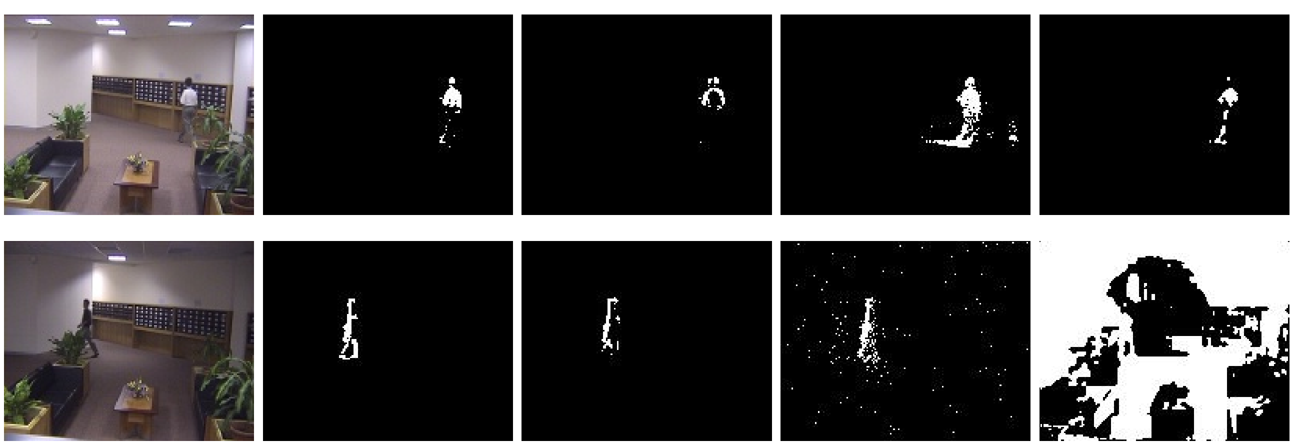


Fig. 4: The results of the *Switchlight* dataset against the four background model: AMM, FDM, GMM and JBFM. First shows the results of the light-on situation and second raw shows the results of the light-off situation.

All the experiments were run on a computer with 8 GB RAM and 2.4 GHz core i7 CPU. We mainly perform three steps: extracting silhouettes using four background models, creating motion descriptors, and classifying data. When modeling the background it takes least time for FDM, followed by AMM. GMM and JBFM take much computational time and it increases when the resolution of the images increases. Although the second step takes much computational time, the average computation time for both training and testing for a dataset of approximately 500 frames is about 1–2 minutes.

VI. Conclusion

In this paper, we presented a technique to detect actions in a frame sequence with dynamic backgrounds. We used both pose characterized by a silhouette and motion in our system by using spatio-temporal descriptors. We used four background models: AMM, FDM, GMM and JBFM. JBFM is better on average recognition rate. Although other simple background models, FDM and AMM, are known to give poor results in the literature, they perform well in some datasets probably due to the motion context descriptor. Computational time for the FDM and AMM is lesser than for other two methods, GMM and JBFM. We carried out several experiments on a diverse set of challenging datasets including Change Detection Challenge Dataset. Our results using JBFM outperform various state-of-the-art algorithms. We were able to detect actions, the difference between "single person walking" and the "two people walking" in the highly dynamic backgrounds by using various test cases. JBFM performs well to detect actions than the other three models. FDM performs poorly in detecting actions. Using these background models our method can detect actions in dynamic backgrounds, similar to action detection in stationary backgrounds.

In this work, we considered only stationary cameras with dynamic backgrounds. Adapting our system to work with moving cameras would also be interesting.

References

[1] X. Ren, T. X. Han, and Z. He, "Ensembe video object cut in highly dynamic scences," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1947–1954.

[2] T. Ko, S. Soatto, and D. Estrin, "Background subtraction on distributions," in *Proceedings of the IEEE European Conference on Computer Vision*, ser. LNCS 5304, vol. Part III. Marseille, France: Springer-Verlag Berlin Heidelberg, 2008, pp. 276–289.

[3] J. Xiao, H. Cheng, H. Sawhney, and F. Han, "Geo-spatial aerial video processing for scene understanding and object tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AL, June 2008, pp. 1 – 8.

[4] ——, "Vehicle detection and tracking in wide field-of-view aerial video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, June 2010, pp. 679 – 684.

[5] A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, Washington, DC, June–July 2004, pp. II–302 – II–309.

[6] J. Rittscher, J. Kato, S. Joga, and A. Blake, "A probabilistic background model for tracking," in *Proceedings of the IEEE European Conference on Computer Vision*, vol. 2, Dublin, Ireland, June-July 2000, pp. 336–350.

[7] A. Elgammal, R. Duraiswami, and L. S. Davis, "Probabilistic tracking in joint feature-spatial spaces," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, Madison, WI, June 2003, pp. 781–788.

[8] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh, "Background modeling and subtraction of dynamic scenes," in *Proceedings of the IEEE International Conference on Computer Vision*, Nice, France, October 2003, pp. 1305–1312.

[9] J. Zhong and S. Sclaroff, "Segmenting foreground objects from a dynamic textured background via a robust Kalman filter," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, Nice, France, October 2003, pp. 44 – 50.

[10] Y. Sheikh and M. Shah, "Bayesian modelling of dynamic scenes for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1778 – 1792, November 2005.

[11] D. Tran, A. Sorokin, and D. Forsyth, "Human activity recognition with metric learning," in *Proceedings of the IEEE European Conference on Computer Vision*, ser. LNCS 5302, vol. Part I. Marseille, France: Springer-Verlag Berlin Heidelberg, 2008, pp. 549–562.

[12] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "changedetection.net: A new change detection benchmark dataset," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2012, pp. 1–8.

[13] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stero vision," in *Proceedings of the International Joint Conferences on Artificial Intelligence*, vol. 2, San Francisco, CA, 1981, pp. 674–679.

[14] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

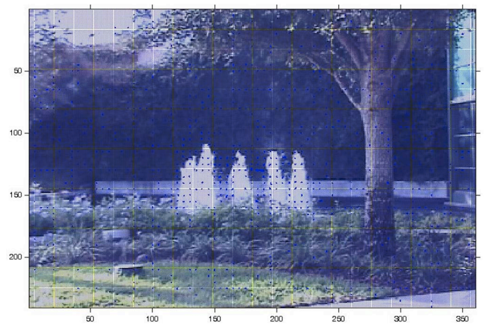
[15] C.-C. Chen and J. K. Aggarwal, "Recognizing human action from a far field of view," in *Proceedings of the IEEE Workshop on Motion and Video Computing*, Snowbird, Utah, December 2009, pp. 1–7.

[16] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proceedings of the IEEE International Conference on Computer Vision*, Corfu, Greece, September 1999.

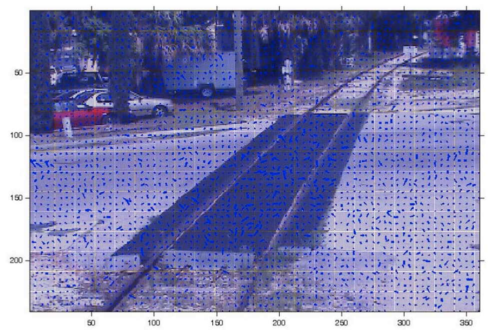
[17] "An abnormal activity datasets," available from http://www.cse.yorku.ca/vision/research/anomalous-behaviour.

[18] L. Li, W. Huang, I. Y. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," in *Proceedings of the ACM International Conference on Multimedia*, 2003, pp. 2–10.

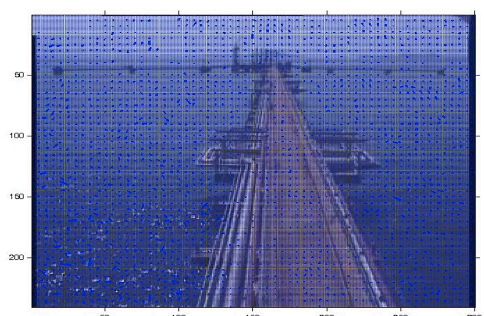
[19] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2012, pp. 38–43.



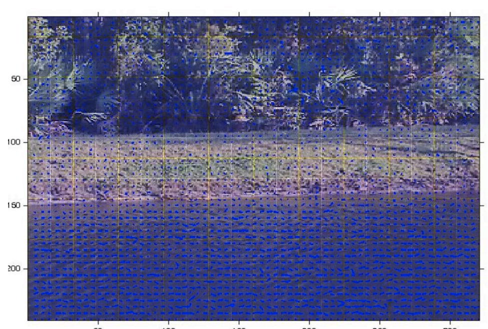
(a)



(b)



(c)



(d)

Fig. 1: Various sources of dynamic behavior. The flow vectors represent the motion in the scene. The arrows indicate motion. (a) The fountain, like the lake-side water, is a temporal texture and does not have exactly-repeating motion (b) a strong breeze can cause nominal motion (camera jitter) of up to 25 pixels between consecutive frames (c) The sea water waves and shimmers (d) The lake-side water ripples and shimmers.

Fig. 3: The datasets are: *Curtain*, *Boat*, *Fountain1*, respectively. First raw are the original images from the respective datasets. Second raw are results of the AMM, third raw are results of the FDM, fourth raw are results of the GMM and finally results of the JBFM.