Moving Object Detection under Object Occlusion Situations in Video Sequences

Dianting Liu, Mei-Ling Shyu, Qiusha Zhu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33146, USA
d.liu4@umiami.edu, shyu@miami.edu, q.zhu2@umiami.edu

Shu-Ching Chen
School of Computing and Information Sciences
Florida International University
Miami, FL 33199, USA
chens@cs.fiu.edu

***Abstract*—It is a great challenge to detect an object that is overlapped or occluded by other objects in images. For moving objects in a video sequence, their movements can bring extra spatio-temporal information of successive frames, which helps object detection, especially for occluded objects. This paper proposes a moving object detection approach for occluded objects in a video sequence with the assist of the SPCPE (Simultaneous Partition and Class Parameter Estimation) unsupervised video segmentation method. Based on the preliminary foreground estimation result from SPCPE and object detection information from the previous frame, an n-steps search (NSS) method is utilized to identify the location of the moving objects, followed by a size-adjustment method that adjusts the bounding boxes of the objects. Several experimental results show that our proposed approach achieves good detection performance under object occlusion situations in serial frames of a video sequence.**

***Keywords*-Moving object detection; Video segmentation; SPCPE; n-steps search**

I. Introduction

Multimedia information plays a more important role than ever in our modern society. Its applications are found in various social fields including advertisements, art, education, entertainment, engineering, medicine, mathematics, business, scientific research, etc. [1][2][3][4]. The explosive amount of multimedia data creates new demands on efficient management, browsing, searching and categorization [5][6][7]. Manual annotation (such as tagging and labeling) cannot catch up the speed of increasing multimedia data, as well as the requirements of fast and automatic organization and searching of the information. These call for the development of efficient and effective methods for automatic multimedia data processing.

As a basic starting point of many relevant topics in multimedia, automatic object detection has attracted many attentions. During the past three decades, the challenge of detecting objects was first presented in processing static images. In natural images, objects seldom laid out in well-separated poses as they often, more or less, overlapped on top of each other. Wittenberg, et al. [8] first clustered neighboring pixels into several regions, yielding a full segmentation of an image, and then combined these regions to objects that carried a semantic meaning. A pixel in an image may be affiliated to one region only, but a region can be part of more than one object. In this way, ambiguities occurred due to overlaps can be resolved on a semantic level. Such an approach was applied to medical images containing overlapping cervical cells, which achieved good results. In [9], the authors presented a new snake algorithm extending conventional snake algorithms by utilizing a pair of stereo images. The authors defined a unique energy function in the disparity space enabling successful boundary detection of the objects even when those objects were overlapped one another and the background was cluttered. An example was presented to demonstrate a successful result of this stereo-snake algorithm for detecting an object out of a complex image, though a set of interested points (including those objects to be segmented) needs to be manually pre-selected. Another novel marker extraction method was proposed to extract markers labeling the target fruit and the background [10]. Based on this marker detection method, a new marker-controlled watershed transform algorithm was developed for accurate contour extraction of the target fruit. The face validity of the segmentation algorithm was tested with a set of grape images, and the segmentation results were overlaid onto the original images for visual inspection. Quantitative comparison was conducted and it showed that the segmentation algorithm can obtain good spatial segmentation results.

With the increasing amounts of digital video data becoming available in the Web, more and more attentions have been paid to content-based video processing approaches that can automatically identify the semantic concepts in a video [11][12][13][14][15]. To achieve this, object detection is a crucial step and thus special attentions are devoted to segmenting a video frame into a set of semantic regions, each of which corresponds to an object that is meaningful to human viewers, such as a car, a person, and a tree. The extra temporal dimension of the video allows the motion of the camera or the scene to be used in processing. Chen, et al. [16] proposed a backtrack-chain-updation split algorithm that can distinguish two separate objects that were overlapped previously. It found the split objects in the current frame and used the information to update the previous frames in a backtrack-chain manner. Thus, the algorithm could

provide more accurate temporal and spatial information of the semantic objects for video indexing. In [17], a region-based spatio-temporal Markov random field (STMRF) model was proposed to segment moving objects semantically and the motion validation was used to detect occluded objects. The STMRF model combined segmentation results of four successive frames and integrated the temporal continuity in the uniform energy function. First, moving objects were extracted by a region-based MRF model between two frames in a frame group of four successive frames. Then, the ultimate semantic object was labeled by minimizing the energy function of the STMRF model. Experimental results showed that their proposed algorithm can accurately extract moving objects.

Some other approaches handled occlusion during object tracking [3][18][19][20][21]. Senior, et al. [19] used appearance models to localize objects during partial occlusions, detect complete occlusions and resolve depth ordering of the objects. The authors reported a good result on the PETS 2001 data set, though the performance was influenced by the pre-selected parameters used to update the probability mask values in the appearance models. [21] maintained a shape prior method to recover the missing object regions during occlusion, while the algorithm was initialized with the boundaries of the objects in the first frame. Stein, et al. [20] proposed a mid-level model for reasoning more globally about object boundaries and propagating such local information to extract improved, extended boundaries with the utilization of subtle motion cues such as parallax induced by a moving camera. The method is mainly a boundary-based algorithm which needs to combine with other techniques to build up a region-based approach for object detection purpose.

From a brief overview of the existing approaches, it shows that many efforts have been made to solve the problem of detecting occluded objects in a video or the sequences of images. However, various kinds of restrictions were imposed before or during the detection processing. For example, domain knowledge was needed in [8][10], interest points [9] or probability parameters [19] needed to be manually pre-selected, the boundaries of the objects in the first frame should be known in [21], etc. Aiming at designing a more generalized detection system, an unsupervised approach is proposed in this paper to identify moving objects under occlusion situations.

This paper is organized as follows. Our proposed framework is presented in Section II, followed by the experimental results and analyses in Section III. We conclude this paper in Section IV.

II. The Proposed Approach

Figure 1 presents the system architecture of the proposed approach for each frame i ($i > 1$) in a video sequence. It includes four steps to handle the occlusion situation between moving objects.

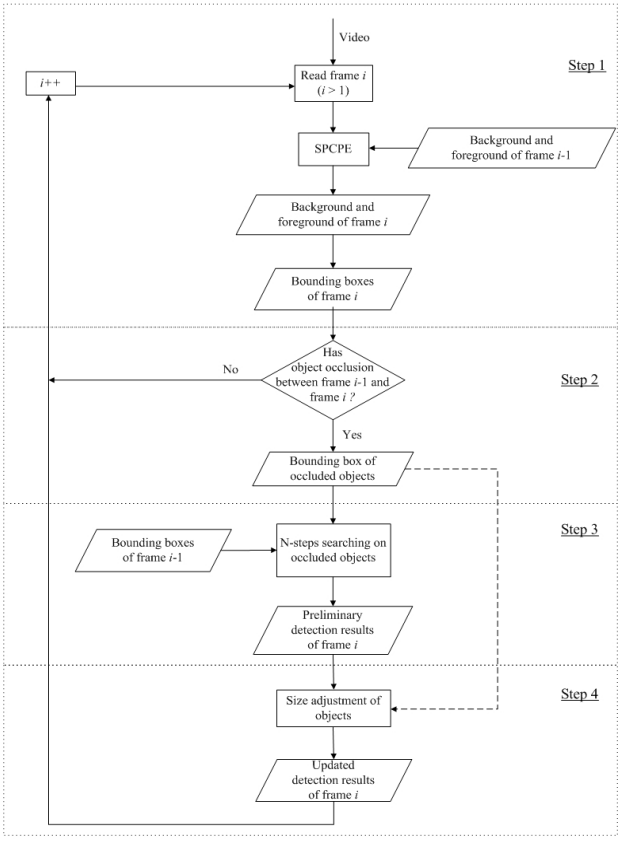


Figure 1. The system architecture of the propose approach

In the first step, background and foreground of frame i are estimated with the help of the unsupervised SPCPE (Simultaneous Partition and Class Parameter Estimation) video segmentation method using the background and foreground of frame i-1 as the initial class partition. After removing the background, bounding boxes are used to describe the foreground objects. Please note that the first frame of the video needs to be processed through SPCPE using an arbitrary initial class partition to get the bounding boxes of its foreground objects. In Step 2, an idea from [16] is adopted to detect object occlusion situations by using the size and location information of bounding boxes in two consecutive frames (i.e., frames i-1 and i). If object occlusion occurs, the bounding box of the occluded objects is passed to Step 3; otherwise, the loop goes back to Step 1 to process the next frame. In order to identify the location of the occluded objects more generally, an n-steps search (NSS) method is employed by using the spatial information of the objects in frame i-1, and the preliminary detection results of frame i are generated in Step 3. Finally, in Step 4, a size-adjustment method is developed to adjust the bounding boxes of the

(a) A video frame

(b) Estimated background and foreground for (a)

Figure 2. An example of the SPCPE estimation result

occluded objects for the purpose of obtaining more accurate sizes and positions of the objects. The same steps are iterated for all the frames i (i >1) of the video.

A. Background and Foreground Estimation

The background and foreground estimation method presented here is based on the SPCPE algorithm [22] that is able to partition objects from the background. The segmentation starts with an arbitrary class partition (for the first frame) and then an iterative process is employed to jointly estimate the class partition and its corresponding class parameters (for the rest of the frames in the video sequence).

The SPCPE algorithm is applied to segment each pixel in frames into two classes, namely background and foreground. Let the segmentation variable be $c = \{c_b, c_f\}$ and the class parameter be $\theta = \{\theta_b, \theta_f\}$. Let all the pixel values y_{ij} (where i and j are the row number and column number of the pixel, respectively) in the frame belonging to class k be put into a vector Y_k, where k = b means background and k = f means foreground. Each row of the matrix Φ is given by $(1, i, j, ij)$, and α_k is the vector of parameters $(\alpha_{k0}, \ldots, \alpha_{k3})^T$.

$$y_{ij} = \alpha_{k0} + \alpha_{k1} i + \alpha_{k2} j + \alpha_{k3} ij, \ \forall (i,j) \ y_{ij} \in c_k \quad (1)$$

$$\mathbf{y}_k = \Phi \alpha_k \quad (2)$$

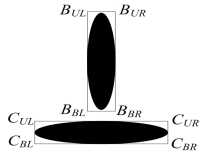
$$\widehat{\alpha}_k = \{\Phi^T \Phi\}^{-1} \Phi^T \mathbf{y}_k \quad (3)$$

Here, it is assumed that the adjacent frames in a video do not differ much, and thus the estimation result of background and foreground of successive frames do not change significantly. Under this assumption, the segmentation of the previous frame is used as an initial class partition, so the number of iterations for processing is greatly decreased. Since the first frame does not have a previous frame, an arbitrary class partition is used to start the estimation process. Figure 2(a) is a color frame extracted from a video of race cars, and Figure 2(b) is its background (shown in black) and foreground (shown in white) estimation result by the SPCPE algorithm.

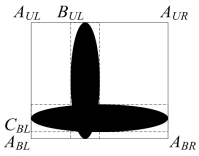
B. Previous Occlusion Detection Strategy

An effective method was proposed in [16] to detect the occlusion of objects by utilizing the concept of minimal bounding rectangle (MBR) in R-trees [23] to bound each semantic object by a rectangle. The main idea is to measure the distances and sizes of the bounding boxes between frames to check if two segments in adjacent frames represent the same object. If a segment cannot find its successor in the subsequent frame, then a merge or split of objects may happen between the two frames.

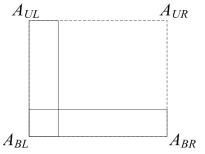
In [16], the authors proposed a backtrack-chain-updation split algorithm and a vertex recovery method to identify the occluded objects, which work well under the situation that two objects with similar sizes and shapes merge or split from the diagonal direction. However, the vertex recovery method may fail in other situations. For example, in Figure 3(a) and Figure 3(b), the vertex recovery method would "paste" vertex B_{UL} onto vertex A_{UL} and vertex C_{BL} onto vertex A_{BL}, leading to the detection result as shown in Figure 3(c), while the correct bounding boxes should be located as shown in Figure 3(d).



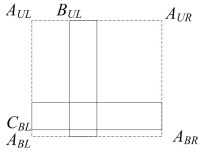
(a) Two objects are separate in a frame



(b) The same objects are occluded in the subsequent frame



(c) Split result by vertex recovery method



(d) Correct detection result

Figure 3. An example when the vertex recovery method would fail

C. New Occluded Objects Detection Approach

Assume that the appearance of the same object in adjacent frames does not change a lot, the idea of a quick block motion estimation method [24], called three-step search (TSS), is extended to identify the location of occluded objects from the spatial information in the previous frame. The TSS algorithm is based on a coarse-to-fine approach with logarithmic decreasing in steps as shown in Figure 4. In TSS, the initial step size is half of the maximum motion displacement p. For each step, nine checking points are matched and the minimum Mean Absolute Difference (MAD) [25] point of that step is chosen as the starting center of the next step whose size is reduced by half. When the step size is reduced to 1, the searching process terminates. The three-step is obviously designed for a small search window (i.e., p = 7).

In this paper, TSS is extended to an n-steps search by using the same searching strategy. For the sake of quick

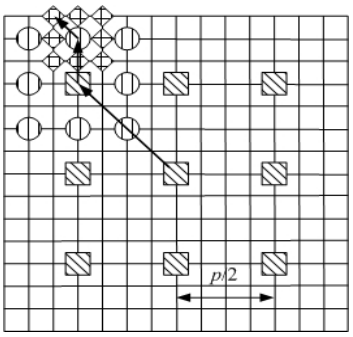


Figure 4. Illustration of three-step search

computation of MAD, the search process is conducted on the SPCPE segmentation result instead of the color frame, and we use the bounding box in the previous frame as the reference block. Figure 5(a) is the SPCPE segmentation result of the current frame where two objects are identified as one. Figure 5(b) is the segmentation result utilizing the positions of the bounding boxes of the previous frame, which is not precise; while on the basis of Figure 5(b), n-steps search returns an acceptable object detection result (as shown in Figure 5(c)).

(a) SPCPE segmentation result in the current frame

(b) Segmentation using bounding boxes of the previous frame

(c) Detection result after n-steps search

(d) Detection result after size adjustment

Figure 5. Detection of occluded objects

D. Size Adjustment of Occluded Objects

The positions of the occluded objects are roughly located by the n-steps search as shown in Figure 5(c). Since the shapes of the moving objects in a video sequence may change, size adjustment is needed to re-size the bounding boxes of the objects. Unlike the size adjustment method in [16] which used the ratio information of size changes on length and width of the split objects in successive frames to update the bounding box of each object, our proposed size adjustment method uses the contour of the occluded objects in the current frame to re-size the object's bounding box.

Table I
THREE EXAMPLE VIDEO SEQUENCES

Category	# of frames	Solution	Source
"Ravens" Video	58	360 * 480	YouTube
"Race cars" Video	51	240 * 320	YouTube
"Girl on the street" Video	81	288 * 352	TRECVid

Let B_{all} denote the bounding box of the occluded objects (shown in Figure 5(a)), and B_{O1} and B_{O2} denote the bounding boxes of the individual objects $O1$ and $O2$ (shown in Figure 5(c)). The final bounding boxes of $O1$ and $O2$ are defined as follows. With the size restriction of the bounding box of the occluded objects, our proposed method has the ability of size adjustment as shown in Figure 5(d).

$$B'_{O1} = B_{O1} \bigcap B_{all}; \tag{4}$$

$$B'_{O2} = B_{O2} \bigcap B_{all}. \tag{5}$$

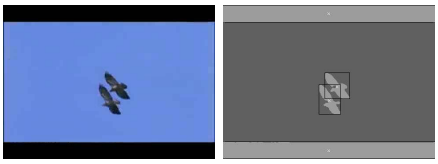
III. EXPERIMENTAL RESULTS AND ANALYSES

Three video sequences containing object occlusion situations are employed to evaluate the performance of the proposed moving object detection approach. Table I lists the information of three video sequences used in our experiments. Two of them are from YouTube [26][27], and the other is from TRECVid 2007 test video collection [28]. Several sample frames in these videos are shown to demonstrate the effectiveness of our proposed approach.

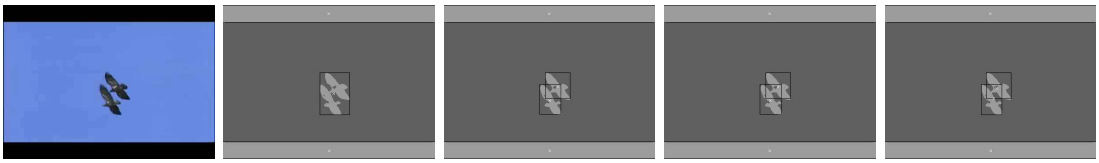
The first column in Figures 6, 7, and 8 shows the original frames from the video sequences. The second column is the segmentation results of the objects from the background by SPCPE. If there are more than two columns, the third column indicates the positions of the bounding boxes of the previous frame, which are used as the initial searching positions of the NSS method on the current frame. The displacement is set to 10 here in the experiment. The searching results are shown in the fourth column, and the fifth column displays the final detection results tuned by the size adjustment method.

Two scenarios are shown in Figure 6. One happens at the beginning of the overlapping of two ravens, and one is the severe occlusion. For the first scenario, the occlusion is not significant, and thus it is easier to get good detection result than in the latter scenario. For the second scenario, one object is heavily occluded by another one, resulting in lots of loss of shape and size information. Therefore, the detection result greatly depends on the previous detection result. It can be seen from the split results in Figure 6(b) and Figure 6(d) that our proposed approach gives a satisfactory performance on both scenarios.

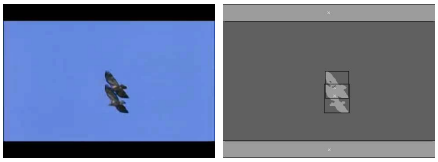
In Figure 7(a), two race cars are segmented into two separate objects by SPCPE; while in Figure 7(b), SPCPE segments them as one partition. Our proposed approach



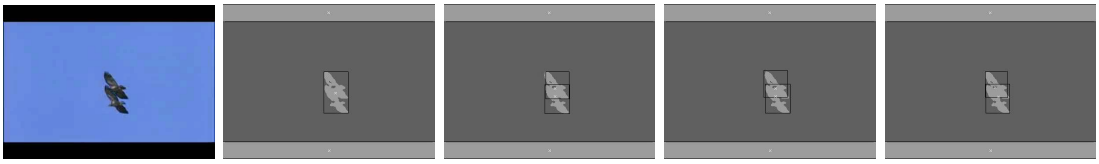
(a) Frame 548



(b) Frame 549



(c) Frame 557



(d) Frame 558

Figure 6. Results for the "Ravens" video sequence

uses the spatial information and the NSS method to roughly detect the new positions of the two cars in Frame 5410. Furthermore, our proposed size adjustment method helps to get more accurate bounding box of each car. Figure 7(c) indicates that when two cars are significantly overlapped, the proposed method is still able to detect their positions correctly. Figure 7(d) shows that after the two cars exchanged their positions, our proposed method still can detect their positions. Figure 7(e) is the last frame (Frame 5428) that the two cars are segmented as one object by SPCPE due to the shadow between the cars, but the detection result on this frame is still very accurate after a series of operations (from Frame 5410 to Frame 5428), which indicates the effectiveness of proposed algorithm.

Figure 8 gives an example that has complicated spatial relationships between objects in a video. In this video, the girl and curb are overlapped and segmented as one partition initially. Figure 8(b) to Figure 8(e) give the detection results. It shows that based on the information from Figure 8(a), our proposed approach successfully splits the girl from the road curb in the following four consecutive frames.

IV. Conclusion

This paper proposes a moving object detection approach utilizing the spatio-temporal information of successive frames in video sequences. It first employs the SPCPE algorithm to estimate the background and foreground of the frames, followed by detecting the object occlusion situations with the help of the size and location information of the bounding boxes in two consecutive frames. Next, the n-steps search and size-adjustment methods are utilized to obtain the preliminary location of each object and tune the size of each object to address the shape changes in the video sequence, respectively. Experimental results on three video sequences with severe object occlusion situations demonstrate that our proposed approach is able to cope with the more generalized object occlusion situations and achieve satisfactory detection results for the moving objects under object occlusion situations.

Acknowledgment

For Shu-Ching Chen, this work is supported in part by NSF HRD-0833093.

References

[1] M. Bae, R. Pan, T. Wu, and A. Badea, "Automated segmentation of mouse brain images using extended MRF," *Neuroimage*, vol. 46, no. 3, pp. 717–725, July 2009.

[2] S.-C. Chen, M.-L. Shyu, M. Chen, and C. Zhang, "A decision tree-based multimodal data mining framework for soccer goal detection," in *IEEE International Conference on Multimedia and Expo (ICME 2004)*, June 2004, pp. 265–268.

[3] S.-C. Chen, M.-L. Shyu, S. Peeta, and C. Zhang, "Spatiotemporal vehicle tracking: The use of unsupervised learning-based segmentation and object tracking," *IEEE Robotics and Automation Magazine, Special Issue on Robotic Technologies Applied to Intelligent Transportation Systems*, vol. 12, no. 1, pp. 50–58, March 2005.

[4] X. Long, W. Cleveland, and Y. Yao, "Multiclass detection of cells in multicontrast composite images," *Computers in Biology and Medicine*, vol. 40, no. 2, pp. 168–178, Feb. 2010.

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, pp. 1–60, 2008.

[6] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of art and challenges," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 2, no. 1, pp. 1–19, 2006.

[7] L. Lin, C. Chen, M.-L. Shyu, and S.-C. Chen, "Weighted subspace filtering and ranking algorithms for video concept retrieval," *IEEE Multimedia*, vol. 18, no. 3, pp. 32–43, 2011.

[8] T. Wittenberg, M. Grobe, C. M*ü*nzenmayer, H. Kuziela, and K. Spinnler, "A semantic approach to segmentation of overlapping objects," *Methods Inf Med*, vol. 43, pp. 343–353, 2004.

[9] S.-H. Kim, J.-H. Choi, H.-B. Kim, and J.-W. Jang, "A new snake algorithm for object segmentation in stereo images," in *IEEE International Conference on Multimedia and Expo*, 2004, pp. 13–16.

[10] Q. Zeng, Y. Miao, C. Liu, and S. Wang, "Algorithm based on marker-controlled watershed transform for overlapping plant unit segmentation," *Optical Engineering*, vol. 48, no. 2, pp. 1–10, 2009.

[11] M. Chen, S.-C. Chen, M.-L. Shyu, and K. Wickramaratna, "Semantic event detection via temporal analysis and multimodal data mining," *IEEE Signal Processing Magazine, Special Issue on Semantic Retrieval of Multimedia*, vol. 23, no. 2, pp. 38–46, October 2006.

[12] A. Hauptmann, M. Christel, and R. Yan, "Video retrieval based on semantic concepts," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 602–622, April 2008.

[13] L. Lin and M.-L. Shyu, "Effective and efficient video high-level semantic retrieval using associations and correlations," *International Journal of Semantic Computing*, vol. 3, no. 4, pp. 421–444, 2009.

[14] Z. Peng, Y. Yang and et al., "PKU-ICST at TRECVID 2009: High level feature extraction and search," in *TRECVID 2009 Workshop*, Nov. 2009.

[15] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Transactions on Multimedia, Special Issue on Multimedia Data Mining*, vol. 10, no. 2, pp. 252–259, Feb. 2008.

[16] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, "Identifying overlapped objects for video indexing and modeling in multimedia database systems," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 4, pp. 715–734, 2001.

[17] W. Zeng and W. Gao, "Semantic object segmentation by a spatio-temporal MRF model," in *The 17th International Conference on Pattern Recognition*, 2004, pp. 775–778.

[18] S.-C. Chen, M.-L. Shyu, S. Peeta, and C. Zhang, "Learning-based spatio-temporal vehicle tracking and indexing for transportation multimedia database systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 4, no. 3, pp. 154–167, September 2003.

[19] A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti, and R. Bolle, "Appearance models for occlusion handling," *Image and Vision Computing*, vol. 24, no. 11, pp. 1233–1243, November 2006.

[20] A. N. Stein and M. Hebert, "Occlusion boundaries from motion: Low-level detection and mid-level reasoning," *Int. J. Comput. Vision*, vol. 82, no. 3, pp. 325–357, May 2009.

[21] A. Yilmaz, X. Li, and M. Shah, "Contour-based object tracking with occlusion handling in video acquired using mobile cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1531–1536, November 2004.

[22] S. Sista and R. L. Kashyap, "Unsupervised video segmentation and object tracking," *Comput. Ind.*, vol. 42, no. 2-3, pp. 127–146, July 2000.

[23] A. Guttman, "R-trees: a dynamic index structure for spatial searching," in *International Conference on Management of Data*, 1984, pp. 47–57.

[24] T. Koga, K. Iinuma, A. Hirano, Y. Iijima, and T. Ishiguro, "Motion-compensated interframe coding for video conferencing," in *Proc. NTC81*, New Orleans, LA., November 1981, pp. C9.6.1–9.6.5.

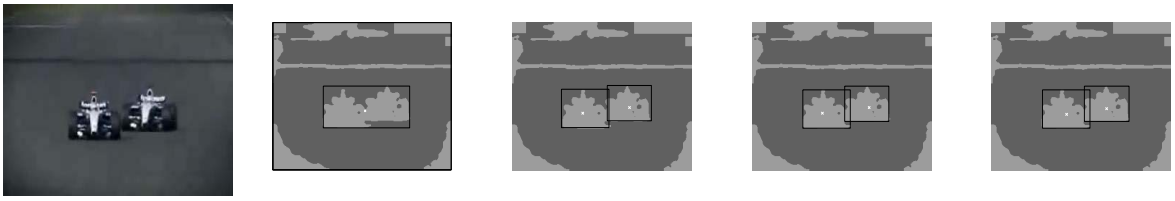
[25] Z.-N. Li and M. S. Drew, *Fundamentals of Multimedia*. Prentice-Hall, 2004.

[26] worldcarfans. (2007, Nov 5) Martin Brundle & Mark Blundell Demonstrating F1 overtaking. [Online]. Available: http://www.youtube.com/watch?v=Koprlc1T4sw&feature=BFa&list=PL80EE1CD57C7842C7&index=29

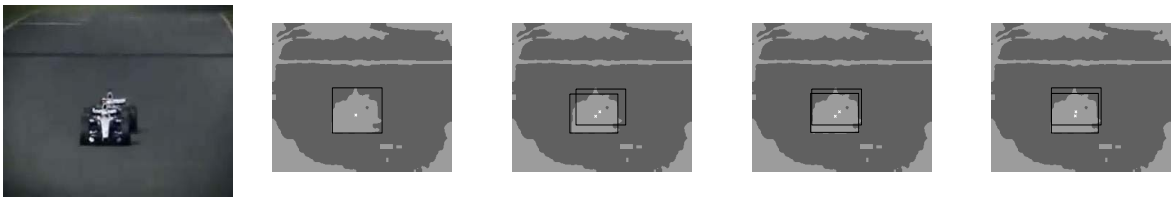
[27] PBS. (2007, Dec 7) Nature, Ravens, Raven Courting Ritual, PBS. [Online]. Available: http://www.youtube.com/watch?v=os5jcMjiXKI&feature=related

[28] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, 2006, pp. 321–330.

(a) Frame 5409

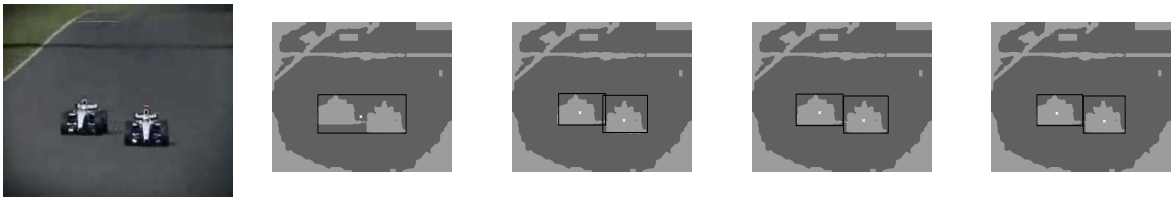


(b) Frame 5410



(c) Frame 5418

(d) Frame 5424

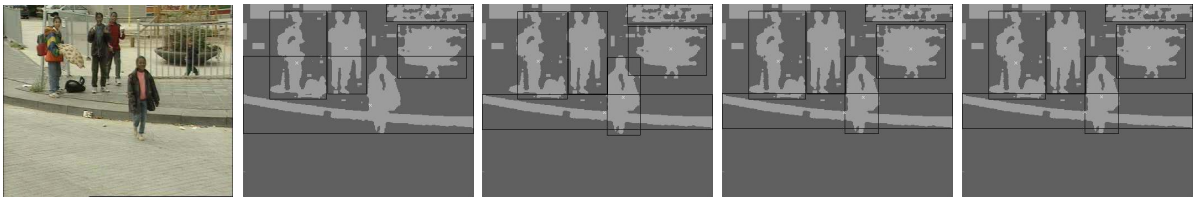


(e) Frame 5428

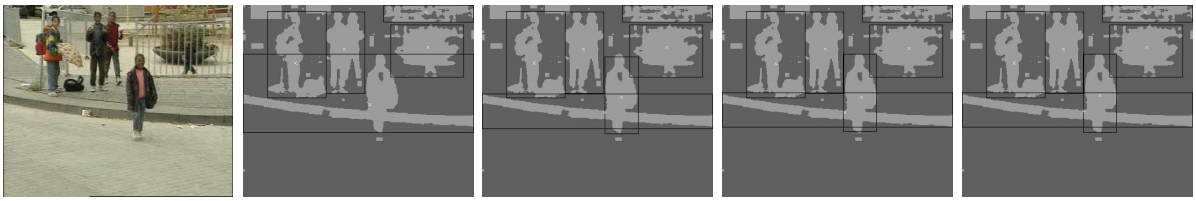
(f) Frame 5429

Figure 7. Results for the “Race cars” video sequence

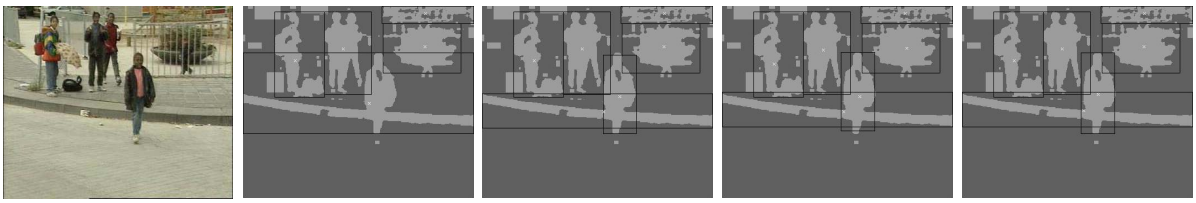
(a) Frame 5284



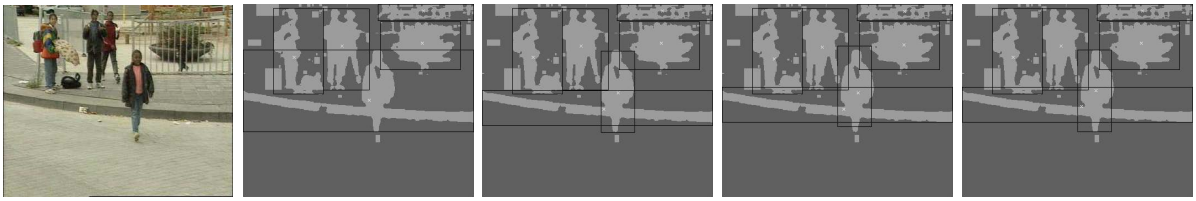
(b) Frame 5285



(c) Frame 5286



(d) Frame 5287



(e) Frame 5288

Figure 8. Results for the “Girl on the street” video sequence