Effective Moving Object Detection and Retrieval via Integrating Spatial-Temporal Multimedia Information

Dianting Liu, Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33124, USA
Email: d.liu4@umiami.edu, shyu@miami.edu

***Abstract*—In the area of multimedia semantic analysis and video retrieval, automatic object detection techniques play an important role. Without the analysis of the object-level features, it is hard to achieve high performance on semantic retrieval. As a branch of object detection study, moving object detection also becomes a hot research field and gets a great amount of progress recently. This paper proposes a moving object detection and retrieval model that integrates the spatial and temporal information in video sequences and uses the proposed integral density method (adopted from the idea of integral images) to quickly identify the motion regions in an unsupervised way. First, key information locations on video frames are achieved as maxima and minima of the result of Difference of Gaussian (DoG) function. On the other hand, a motion map of adjacent frames is obtained from the diversity of the outcomes from Simultaneous Partition and Class Parameter Estimation (SPCPE) framework. The motion map filters key information locations into key motion locations (KMLs) where the existence of moving objects is implied. Besides showing the motion zones, the motion map also indicates the motion direction which guides the proposed "integral density" approach to quickly and accurately locate the motion regions. The detection results are not only illustrated visually, but also verified by the promising experimental results which show the concept retrieval performance can be improved by integrating the global and local visual information.**

***Keywords*-Spatial-temporal; moving object; key motion location; integral image; integral density; SPCPE.**

I. Introduction

With the rapid advances of Internet and Web 2.0, the amount of online multimedia data increases in an explosive speed, which brings many challenges to data retrieval, browsing, searching and categorization [1][2][3]. Manual annotation obviously cannot catch up the speed of increasing multimedia data, so content-based video processing approaches are developed to quickly and automatically identify the semantic concepts and annotate the video sequences [4][5][6][7][8].

Automatic object detection techniques, as a key step in content-based multimedia data analysis framework, has also attracted lots of attention these years. It aims to segment a visual frame into a set of semantic regions, each of which corresponds to an object that is meaningful to the human vision system, such as a car, a person, and a tree. When the object detection issues move from image area to video domain, temporal information in video sequences brings moving object-level information which can be utilized for moving object detection. From this perspective, this paper integrates spatial information locations (the yellow crosses shown in Fig. 1(a)) and temporal motion cues (the white and black zones shown in Fig. 1(b)) to find the locations that are rich in spatial-temporal information (the yellow crosses shown in Fig. 1(c)) and use integral density method to identify the motion region (the yellow bounding box shown in Fig. 1(d)). The motion region is also verified to be helpful to improve the content-based multimedia retrieval performance.

Psychological studies find that a human vision system perceives external features separately [9] and is sensitive to the difference between the target region and its neighborhood. Such kind of high contrast is more likely to attract human's first sight than their surrounding neighbors [10]. Following this finding, many approaches have focused on the detection of feature contrasts to trigger human vision nerves. This research field is usually called visual attention detection or salient object detection. Liu, et al. [11] employed a conditional random field method which is learned to effectively combine multiple features (including multi-scale contrast, center-surround histogram, and color spatial distribution) for salient object detection.

In video sequences, the action of objects will dominate the frame and human perceptual reactions will mainly focus on motion contrast regardless of visual texture in the scene. Several researchers have extended the study from the spatial attention to the temporal domain where prominent motion plays an important role. Chen, et al. [12] 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. Liu, et al. [13] extended Chen's work to process the more generalized overlapped situation. In [14], the authors proposed a spatiotemporal video atten-

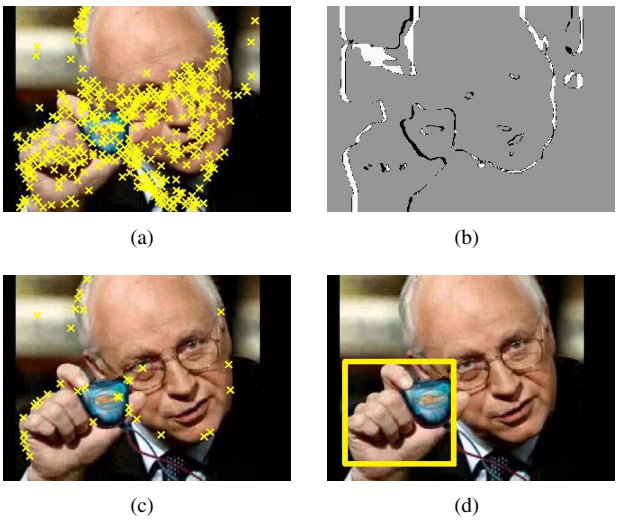


Figure 1. Illustration of moving object detection. (a) Key information locations (yellow crosses) on a sample image. (b) Motion map of the sample image (white and black zones). (c) Key motion locations (KMLs) on the motion map. (d) Motion region detected by the proposed integral density method from (c).

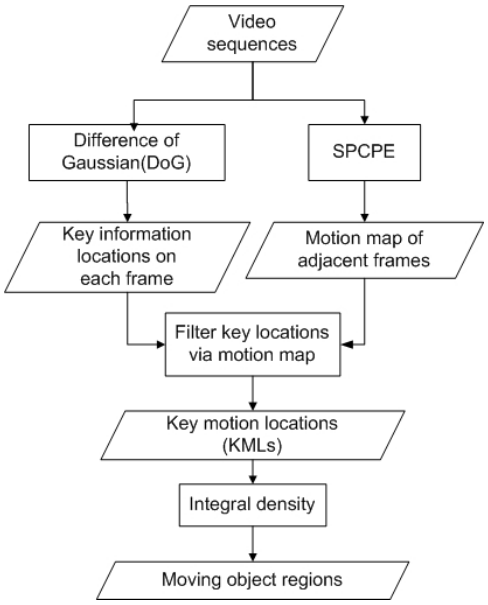


Figure 2. The proposed moving object detection framework

tion detection technique for detecting the attended regions that correspond to both interesting objects and actions in video sequences. The presented temporal attention model in the paper utilized the interest point correspondences (instead of the traditional dense optical fields) and the geometric transformations between images. Motion contrast was estimated by applying RANSAC (RANdom SAmple Consensus) on point correspondences in the scene. Obviously, the performance of the temporal attention model is greatly influenced by the results of point correspondences.

The main contributions of this paper include: (1) Define a motion map based on the segmentation results of Simultaneous Partition and Class Parameter Estimation (SPCPE) [15] and identify key motion locations (KMLs) by filtering key information locations via the motion map. The motion map not only shows the motion areas, but also indicates the moving direction of the objects which help the identification of the moving objects later. (2) Propose an integral density method inspired by the idea of integral image in order to quickly and accurately detect the moving object regions from KMLs. (3) Present a multimedia retrieval framework to integrate global and local features in order to enhance the existing retrieval framework that uses only global features.

The remainder of this paper is organized as follows. The moving object detection framework is presented in Section 2. Section 3 describes the proposed moving object detection and retrieval model that fuses the global and local features to enhance the retrieval performance. The new content-based multimedia retrieval framework is also introduced in this section. Section 4 presents the experimental results and analyzes the performance from the detection and retrieval angles, respectively. Section 5 concludes the proposed moving object detection and retrieval model.

II. Moving Object Detection Framework

In the motion detection field, the optical flow method is commonly used to compute motion contrast between visual pixels. However, it has obvious drawbacks. For instance, when multiple motion layers exist in the scene, optical flows at the edge pixels are noisy. Also, in texture-less regions, optical flows may return error values. To address these drawbacks, instead of using the above pixel-wise computations, we employ an unsupervised object segmentation method called SPCPE (Simultaneous Partition and Class Parameter Estimation) to segment the frame approximately, and then compute the difference between the two frame segments whose results are called the "motion map" in this paper. This motion information is used to filter the key information locations obtained from the result of difference of Gaussian (DoG) function applied in scale space to a series of smoothed and re-sampled videos frames [16]. Finally, the integral density method is utilized to identify those regions as the moving objects where the density of key motion locations (KMLs) is high. Fig. 2 illustrates the process of our proposed moving object detection framework.

A. Motion map generation

We aim to separate the moving objects from the relatively static background in an unsupervised manner or a bottom-up approach. Unlike those top-down approaches which are task-driven and need to know the prior knowledge of the target, bottom-up approaches are referred to as the stimuli-driven mechanism which is based on the human reaction

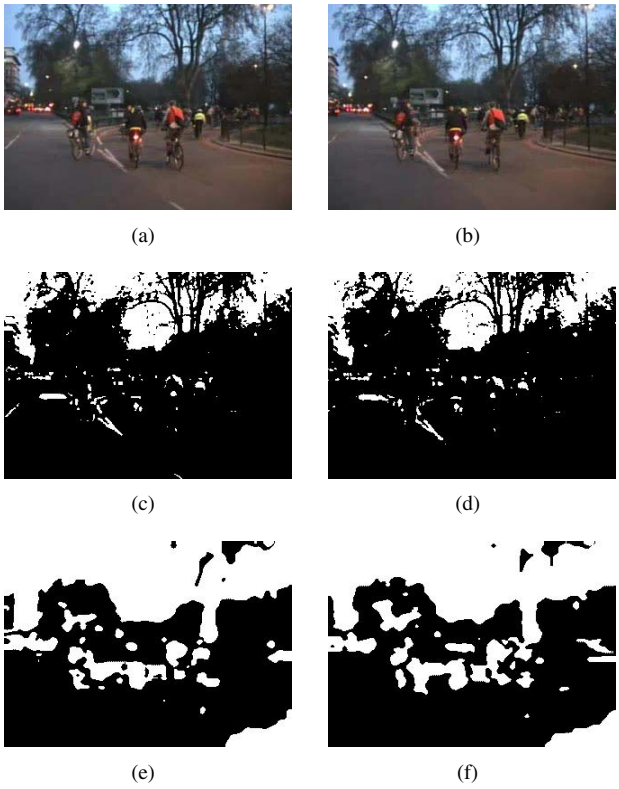


Figure 3. Comparison of the binary images and SPCPE segmentation results of sample frames. (a) and (b) are adjacent frames containing trees, roads, and bicycling persons. (c) and (d) are binary images converted from (a) and (b). (e) and (f) are the two-class SPCPE segmentation results of (a) and (b).

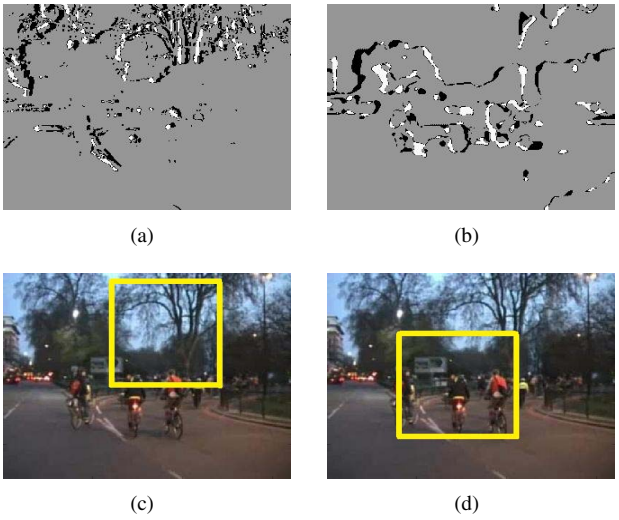


Figure 4. Comparison of the motion maps and corresponding moving object detection results. (a) is the motion map generated from Fig. 3(c) and Fig. 3(d). (b) is the motion map created from Fig. 3(e) and Fig. 3(f). (c) and (d) are detection results by using (a) and (b).

to the external stimuli (for example, the prominent motion from the surroundings).

As shown in the third row of Fig. 3, the pixels in the video frames are segmented into two classes by using the SPCPE algorithm. It starts with an arbitrary class partition and then an iterative process is employed to jointly estimate the class partition and its corresponding class parameters. The iteration is terminated when the areas of the two classes are stable. Assume that the content of the adjacent frames in a video sequence does not change much (as shown in Fig. 3(a) and Fig. 3(b)), and thus the estimation result of the two classes of successive frames does not differ a lot as shown in Fig. 3(e) and Fig. 3(f). Under this assumption, the segmentation of the previous frame is used as an initial class partition for the next frame, so the number of iterations for processing is significantly decreased.

Though the contours of the objects are not very precise as shown in Fig. 3(e) and Fig. 3(f), the segmentation is considered to reflect the object information in the frame. Even though using binary images as shown in Fig. 3(c) and Fig. 3(d) can in some degrees represent object information, the difference of binary images shown in Fig. 4(a) contains too much noise so that it fails to give the motion cues of moving objects as the difference of the SPCPE results does in Fig. 4(b). Assume the white regions and black regions in Fig. 3(e) and Fig. 3(f) stand for class 1 and class 2, respectively. The gray area in Fig. 4(b) shows the pixels which do not change the class labels from Fig. 3(e) to Fig. 3(f). The white zones in Fig. 4(b) shows those pixels which change from class 1 to class 2, and the black zones shows those pixels which change from class 2 to class 1. Obviously, these white and black zones contain the contour information of the moving objects and background, as well as the moving direction information of the objects. Thus we define the white and black zones in Fig. 4(b) as the motion map of Fig. 3(e) and Fig. 3(f). Fig. 4(c) and Fig. 4(d) are the motion detection results by using different motion maps (Fig. 4(a) and Fig. 4(b)), respectively. It shows that SPCPE aims to keep the general object information while ignoring the detailed texture information, so it is good for getting a robust motion map. In contrast, the binary images contain more detailed object contour information which may influence the quality of the motion map if the background in the frames contains many detailed texture information.

B. Key motion locations identification via motion map

Our proposed moving object detection and retrieval model identifies the key information locations by searching the maxima and minima of the results of the DoG function when it is applied in scale space to a series of smoothed and re-sampled frames [16]. Some of the key information locations describe the moving objects and the others describe the background. Based on this observation, we use the motion map generated in the previous step to filter those key information locations which are not located on the contour

of the moving object. Actually, only the key information locations on the motion map are kept as the so-called "key motion locations" (KMLs) to help find the moving object regions as shown in Fig. 1(c), since we consider KMLs are motion related.

C. Moving object region detection

After identifying KMLs, how to group them into meaningful moving objects becomes a critical issue. This is a global searching problem that is very time-consuming. To solve this problem, we propose a method to quickly find the moving object regions that have a high density of KMLs and satisfy the direction constraint in the motion map. In the proposed model, the idea of integral images, as defined in [17], is adopted since it allows the fast implementation of the box type convolution filters. The entry of an integral image $I_{\sum}(\mathbf{x})$ at a location $\mathbf{x}=(x,y)$ represents the sum of all pixels in the input image I of a rectangular region formed by the point $\mathbf{x}$ and the origin, i.e.,

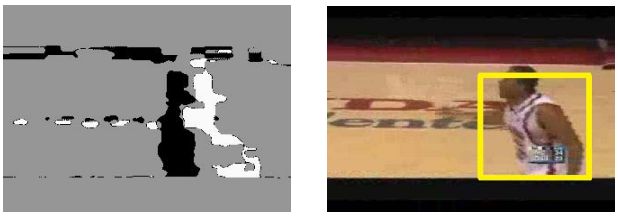
$$I_{\sum}(\mathbf{x}) = \sum_{i=0}^{i\leq x} \sum_{j=0}^{j\leq y} I(i,j).$$

With $I_{\sum}(\mathbf{x})$ calculated, it only takes four additions to calculate the sum of the intensities over any upright rectangular areas, independent of their sizes.

Inspired by the integral images, we calculate the density of KMLs in the input image instead of the sum of all pixels. This new approach is defined as the "integral density" in this paper. This provides us a fast way to find the region where the density of KMLs is high, and we consider this region is greatly related to the moving objects. In order to bound the whole moving object instead of part of it, the satisfied region is subject to one condition. That is, the moving object region needs to satisfy the constraint that the ratio of two motion zones (white zone and black zone) in the motion map is not high. Ideally, the two zones should have the same area, which indicates the moving direction of the object. Please note that in the paper, the ratio is set to 2. However, the determination of this ratio threshold can be adjusted depending on the applications. As shown in Fig. 5(a), the white zone and black zone are separate. Without the above constraint, only a half of the person in Fig. 5(b) may be bounded where the density of the interest points is high.

III. Effective Retrieval Using Global and Local Features

Our proposed moving object detection and retrieval model consists of a new content-based multimedia retrieval framework that integrates the global and local features to enhance the retrieval performance. The motivation of this framework is to utilize the information obtained from the moving object detection part of the model so that the local or object-level features can be integrated with the commonly used global features for the retrieval. As shown in Fig. 6, the training phase of the retrieval framework includes two main modules:



(a) (b)

Figure 5. Demonstrate the necessity of the proposed constraint for integral density. (a) is the motion map of frame (b), which shows a correct moving object detection result under the constraint.

feature extraction and subspace training, which work on the moving object regions and original frames, respectively. The representative subspace projection modeling (RSPM) algorithm [8] is adopted to train the subspace in this proposed content-based multimedia retrieval framework. That is, a subspace called local subspace will be trained for the local features extracted from the moving object regions, and a subspace called global subspace will be trained for the global features extracted from the original video frames.

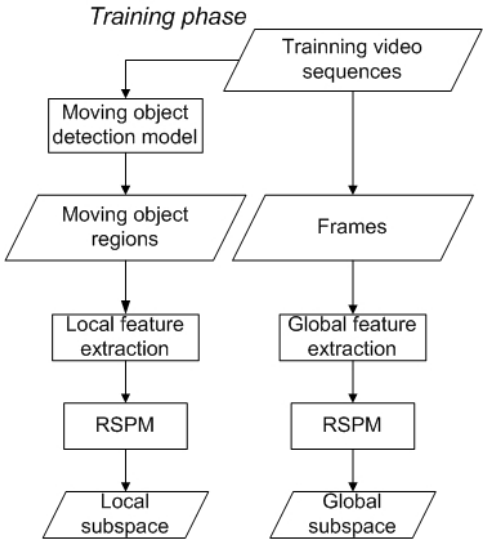


Figure 6. Training phase of the proposed moving object detection and retrieval model (with a new content-based multimedia retrieval framework)

For the testing phase that is shown in Fig. 7, the feature extraction process is the same as that in the training phase. The visual features are projected onto the subspace obtained in the training phase. That is, the local features extracted from the moving object regions in the testing data set will be projected onto the local subspace obtained in the training phase (from the moving object regions in the training data set), and the global features extracted from the video frames in the testing data set will be projected onto the global subspace obtained in the training phase (from the video frames in the training data set). Each testing feature vector

will be converted into a similarity score after the subspace projection. A fusion process is necessary to combine the similarity scores from local and global subspaces to give a final similarity score to represent each video shot. In this paper, the logistic regression method is employed to fuse the global and local similarity scores from the different features. In the future, other fusion methods will be explored in our proposed model.

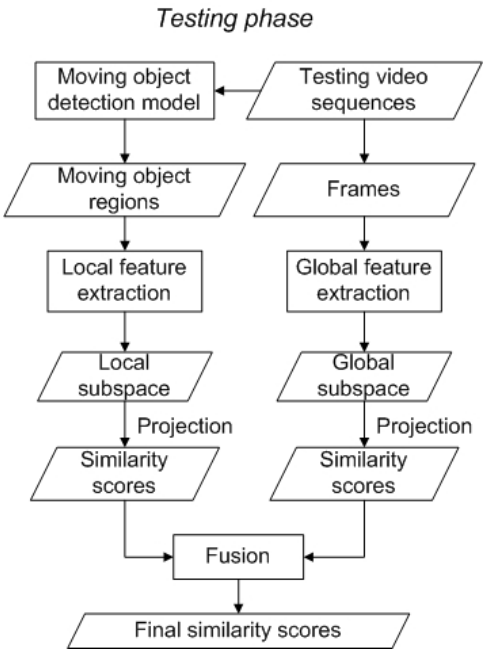


Figure 7. Testing phase of the proposed moving object detection and retrieval model (with a new content-based multimedia retrieval framework)

IV. Experimental results and analyses

The effectiveness of our proposed moving object detection and retrieval model is evaluated on a subset of the TRECVID 2010 video collection [18]. The subset contains ten queries (concepts) which involve motion (consisting of motion information) as shown in Table I. Some shots are multi-labeled, so the total number of shots is less than the sum of the numbers of shots in all ten queries. For example, a shot which is annotated as "running" is possibly also labeled as "sports".

A. Performance of the moving object detection framework

TRECVID 2010 video collection provides a reference key frame for each shot. Assume the reference key frame stands for the content of the shot, the proposed moving object detection and retrieval model also extracts four extra frames around the reference key frame in each shot for the purpose of calculating the motion of the shots. In the experiments, the time interval between two frames is set to 0.2 seconds. This value will be adaptively computed based on the motion speed in the shot in our future work. Also, for the fast computation purpose, the minimum motion region size is set to 0.4 times of the shorter dimension size of the frame based on the assumption that a small region only includes a part of a moving object. Fig. 8 shows some moving object detection and retrieval examples.

Table I
Data set parameters of the experiments

Concept ID	Concept	# of shots in training	# of shots in testing
4	Airplane-flying	83	113
6	Animal	687	1069
13	Bicycling	79	55
38	Dancing	390	250
59	Hand	759	287
100	Running	245	116
107	Sitting down	1555	536
111	Sports	607	839
127	Walking	1067	412
128	Walking or running	2145	766
	Total	7617	3604

Please note that the key information location distribution in the first column represents a rich texture information area from the spatial angle and does not consider the temporal motion information in the video sequences. As the result of filtering the key information locations via the motion map in the third column obtained from the SPCPE algorithm, the remaining key information locations are updated to the key motion locations (KMLs) and keep the spatial-temporal information which is suitable for the moving object detection purposes as shown in the fourth column of Fig. 8. The last column of Fig. 8 is the detection results from the temporal model of [14]. The reason why the model proposed in [14] fails in some cases is that the model is greatly influenced by the results of point correspondences. Though the new model proposed in this paper uses a similar strategy to locate the key information locations as [14], the proposed motion map removes those motion-unrelated information and the integral density method successfully gets the moving object region by precisely analyzing the distribution of the KMLs. Fig. 8 is an example that illustrates the effectiveness of the proposed moving object detection model.

B. Performance of the proposed moving object detection and retrieval model

The motion regions detected in the previous subsection can be viewed as a kind of local features that describe the object-level texture of the shot. This may be complementary to the global information for multimedia retrieval. To verify this assumption, a set of comparable experiments is conducted in three data sets (reference key frame (RKF), multiple frames (MF), and multiple frames plus motion regions (MF+MR)). The data set of the reference key frames is the same set used in the moving object detection model. Four frames are extracted per shot around the reference

Table II
COMPARISON OF MAP FOR DIFFERENT NUMBERS OF RETRIEVAL SHOTS BETWEEN THREE DATA SETS USING YCBCR FEATURES

Top	10	100	1000	2000	All
RKF	0.4528	0.4657	0.3745	0.3400	0.3063
MF	0.5484	0.5503	0.4147	0.3752	0.3460
MF+MR	**0.6746**	**0.6103**	**0.4355**	**0.3968**	**0.3691**

Table III
COMPARISON OF MAP FOR DIFFERENT NUMBERS OF RETRIEVAL SHOTS BETWEEN THREE DATA SETS USING GABOR FEATURES

Top	10	100	1000	2000	All
RKF	0.7034	0.5773	0.4045	0.3630	0.3349
MF	0.6511	0.6051	0.4563	0.4051	0.3798
MF+MR	**0.7286**	**0.6550**	**0.4799**	**0.4271**	**0.3997**

key frame. That is, including the reference key frame, each shot is represented by five frames which consist of the MF data set. On each frame, one or more motion regions (MR) may be detected. Motion regions detected in the MF data set plus the MF data set itself form the MF+MR data set. The experimental design aims to check whether the motion region features can complement the global features to enhance multimedia retrieval.

In the feature extraction step, three kinds of texture features (YCbCr, Gabor, and LBP) are extracted from each data set. For YCbCr features, the frame or region is first converted to the YCbCr color space from the RGB color space. Then the frame or region is divided into nine blocks. Mean, variance, skewness, and kurtosis are calculated on Y, Cb, and Cr components, respectively. Considering the mean, variance, skewness, and kurtosis calculated on Y, Cb, and Cr components of the global frame, there are totally 120 features that are obtained from each frame or region. For Gabor features, a set of Gabor filters with different frequencies and orientations is convolved with the frame or region to generate 108 features to describe the frame or region. LBP (Local Binary Pattern) is a simple yet very efficient texture operator which labels the pixels of a frame or region by thresholding the neighborhood of each pixel and considers the result as a binary number. After summarization of the binary numbers, 59 LBP features are returned to represent the frame or region.

Table IV
COMPARISON OF MAP FOR DIFFERENT NUMBERS OF RETRIEVAL SHOTS BETWEEN THREE DATA SETS USING LBP FEATURES

Top	10	100	1000	2000	All
RKF	0.4929	0.5287	0.4370	0.4079	0.3915
MF	0.5316	0.5541	0.4729	0.4437	0.4281
MF+MR	**0.6209**	**0.6034**	**0.4983**	**0.4659**	**0.4501**

Table V
COMPARISON OF MAP FOR DIFFERENT NUMBERS OF RETRIEVAL SHOTS BETWEEN THREE DATA SETS USING ALL THREE KINDS OF FEATURES

Top	10	100	1000	2000	All
RKF	0.7159	0.6787	0.5460	0.5113	0.4952
MF	0.7801	0.7221	0.5759	0.5423	0.5261
MF+MR	**0.8563**	**0.7741**	**0.6134**	**0.5748**	**0.5594**

In this paper, we transform the multi-class issue into the binary class problem. This means that in the training phase, the one-again-all strategy is utilized. Logistic regression method is used to fuse the similarity scores of multiple frames (MF) as well as multiple frames and motion region (MF+MR).

The criterion used in the paper to evaluate the performance of different approaches is the mean average precision (MAP) which is the mean of the average precision for each query. Average precision (AP) is a popular measure that takes into account both recall and precision in the information retrieval field. Strictly speaking, the average precision is the precision averaged across all values of the recall between 0 and 1. In practice, the integral is closely approximated by a sum over the precisions at every possible threshold value, multiplied by the change in recall.

Tables II, III, IV, and V show the MAP values when retrieving 10, 100, 1000, 2000, and all shots in the three data sets. In these tables, RKF means the reference key frame data set; MF means the multiple-frame data set including the reference key frame and four extra frames; and MF+MR is the union of MF and motion-region data set, including multiple frames with the moving object region obtained from the multiple frames. Though using different features, the retrieval results are consistent among three kinds of data sets. The results of MF generally outperform those of RKF at different numbers of the retrieval shots, which indicates that using multiple frames could provide more useful information to improve the concept retrieval performance than using a single reference key frame. On the other hand, MR+MF outperforms both MF and RKF on all ten queries. This verifies that the moving object region has the concept-related information that can be utilized in the semantic retrieval domain. When comparing the MAP values in the same data set among Tables II, III, and IV, the YCbCr, Gabor, and LBP return similar MAP values. If using multiple kinds of features, the retrieval performance would be improved in a considerable degree (20% more in Table V). Also, we observe that the proposed moving object detection model indeed effectively identifies the moving object in the frame as shown in Fig. 8.

V. CONCLUSION

This paper proposes a new moving object detection and retrieval model to analyze and retrieve the spatial-temporal video sequence information. A motion map is generated from the SPCPE segmentation results to keep the motion related key information locations, called key motion locations (KMLs). Next, an integral density method is proposed to quickly and precisely identify the motion region by analyzing the density of the KMLs under the motion direction

restraint generated by the motion map. A new multimedia retrieval framework using the global and local features are presented to effectively combine and fuse the texture information from the global features via the original frames and the local features from the motion regions. Experimental results show that our proposed moving object detection and retrieval model achieves good performance in terms of the moving object detection and multimedia concept retrieval.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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, December 2009.

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

[8] 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, February 2008.

[9] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, January 1980.

[10] J. Duncan and G. Humphreys, "Visual search and stimulus similarity," *Psychological Review*, vol. 96, no. 3, pp. 433–58, July 1989.

[11] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 33, no. 2, pp. 353–367, February 2011.

[12] 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.

[13] D. Liu, M.-L. Shyu, Q. Zhu, and S.-C. Chen, "Moving object detection under object occlusion situations in video sequences," in *Proceedings of the 2011 IEEE International Symposium on Multimedia*, 2011, pp. 271–278.

[14] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proceedings of the 14th annual ACM international conference on Multimedia*, 2006, pp. 815–824.

[15] S. Sista and R. L. Kashyap, "Unsupervised video segmentation and object tracking," *Computers in Industry Journal*, vol. 42, no. 2-3, pp. 127–146, July 2000.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal on Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[17] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[18] 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.

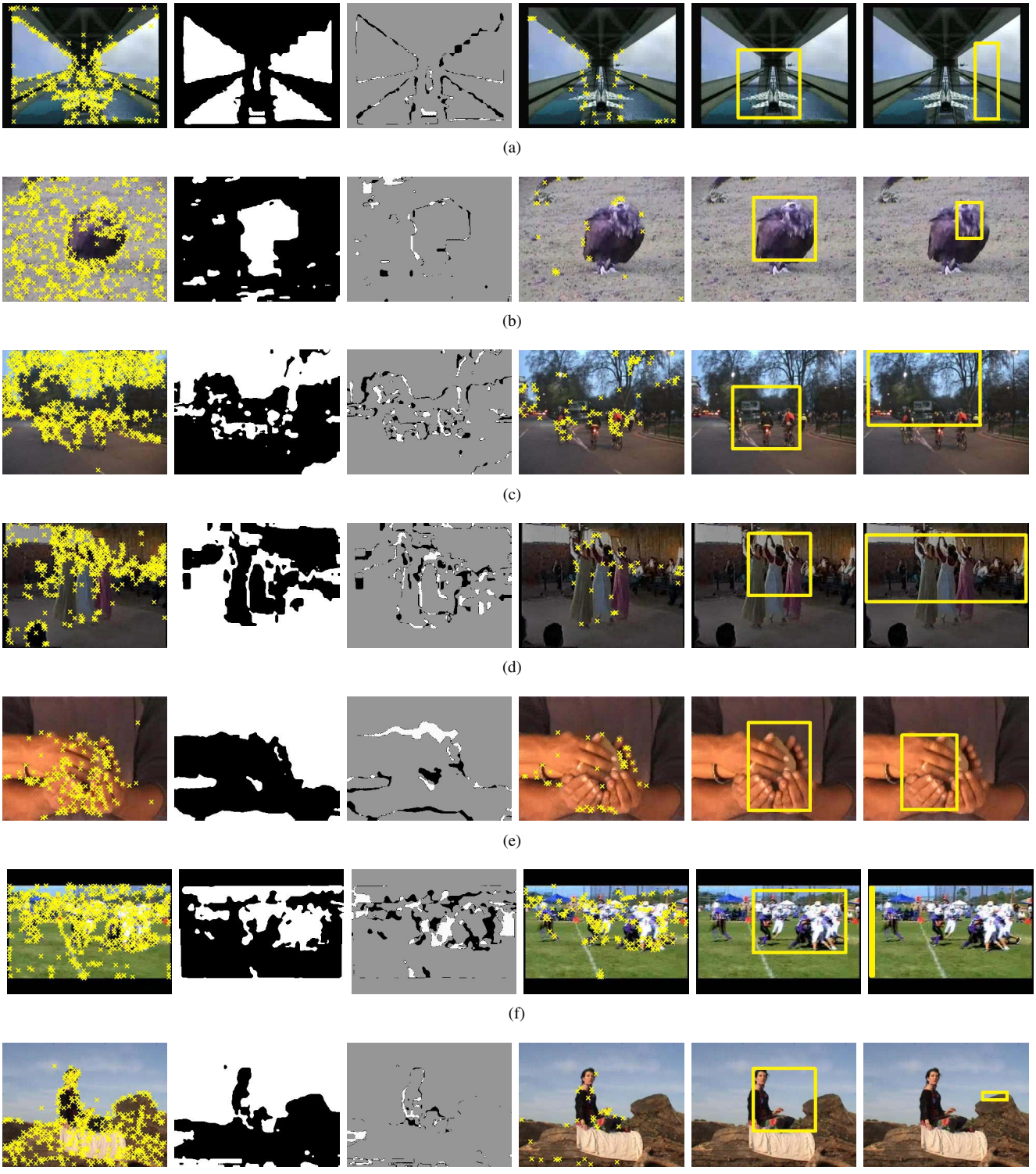


Figure 8. Sample moving object detection and retrieval results compared with [14]. From left to right: key information locations, SPCPE segmentation results, the motion map, key motion locations (KMLs), moving object regions obtained by the proposed model, and object regions obtained by temporal model in [14]. From top to bottom: example images of (a) airplane-flying, (b) animal, (c) bicycling, (d) dancing, (e) hand, (f) running, (g) sitting down.