Semantic Retrieval for Videos in Non-Static Background Using Motion Saliency and Global Features

Dianting Liu and 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 this paper, a video semantic retrieval framework is proposed based on a novel unsupervised motion region detection algorithm which works reasonably well with dynamic background and camera motion. The proposed framework is inspired by biological mechanisms of human vision which make motion salience (defined as attention due to motion) is more "attractive" than some other low-level visual features to people while watching videos. Under this biological observation, motion vectors in frame sequences are calculated using the optical flow algorithm to estimate the movement of a block from one frame to another. Next, a center-surround coherency evaluation model is proposed to compute the local motion saliency in a completely unsupervised manner. The integral density algorithm is employed to search the globally optimal solution of the minimum coherency region as the motion region which is then integrated into the video semantic retrieval framework to enhance the performance of video semantic analysis and understanding. Our proposed framework is evaluated using video sequences in non-static background, and the promising experimental results reveal that the semantic retrieval performance can be improved by integrating the global texture and local motion information.**

***Keywords*-Video semantic retrieval, motion saliency, motion detection, global feature, non-static background.**

I. Introduction

In the latest two decades, semantic retrieval on text and multimedia content has become an important research field [1][2]. A number of text-based search engines appear in the Internet for topic and event searching. Although the text data shares certain correlation information with content data, searching for a multimedia content is not as easy because multimedia data, as opposed to text data, needs more steps of pre-processing to yield indices relevant for querying [3][4]. On the other hand, multimedia data (e.g., an image or a video sequence) can be interpreted in many ways and there is no commonly agreed-upon vocabulary [5][6][7]. Thus, for a large multimedia database, the traditional way of manually assigning a set of labels to a record, storing it and matching the stored label with a query obviously is not feasible and effective. Specially, the rapid advances of Internet and Web 2.0 make the amount of online multimedia data increase in an explosive speed, which brings many challenges to data searching, categorization, retrieval, and browsing [8][9][10]. Manual annotations obviously cannot catch up the speed of the increasing multimedia data. Recent research has focused on the use of semantic features of images and videos to automatically index or retrieve the multimedia data [11][12][13][14].

Object-level information extraction is a key step in multimedia semantic analysis frameworks and has attracted broad attention these years. The common way is 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. After years of development, a batch of object detection models have been proposed and achieved good performance in videos captured in controlled backgrounds. However, the uncontrolled videos are the major challenges to the multimedia retrieval engines on the Internet which call for rapid summarization and processing algorithms, including videos recoded by an amateur using a hand-held camera containing significant camera motion, background clutter, and changes in object appearance, scale, and illumination conditions. The most existing techniques of object detection have the requirements of either static cameras or approximate compensation of camera motion; otherwise, they need the foreground objects to move in a consistent direction or have faster variations in appearance than the background. Of course, if explicit background models are available, many methods can get satisfied results [15]. However, the background learning requires a training set of background-only images [16] or batch processing (e.g., median filtering [17]) of a large number of video frames, which must be repeated for each scene and is difficult for dynamic scenes (where the background changes continuously). Toward model robustness, the above requirements are mostly unrealistic and particularly questionable when an ego-motion happens, e.g. a camera that tracks a moving object in a manner such that the latter has very small optical flow, or the background is dynamic.

On the other hand, neurophysiological experiments on primates have shown that neurons in the middle temporal visual area compute local motion contrast with center-surround mechanisms. It has, in fact, been hypothesized that such neurons underlie the perception of motion pop-out and figure-ground segmentation [18]. This evidence suggests that spatio-temporal saliency or foreground motion detection

978-0-7695-5119-7/19 $26.00 © 5119 IEEE
DOI 10.1109/ICSC.2013.57

(a) Airplane-flying (b) Animal

(c) Bicycling (d) Dancing

(e) Hand (f) Sports

(g) Running (h) Walking

Figure 1. Reference key frame examples extracted from the TRECVID 2010 video database of eight concepts with 9130 video shots.

techniques which 1) rely on grouping of features by motion similarity to identify foreground objects or 2) require compensation of camera motion, will have difficulties to match the performance of biological systems. To mimic the human vision system, a center-surround coherency model is proposed in the paper to address the detection limitations of the salient motions under non-static background scenarios. Saliency is defined that there is a region in a scene that is more "attractive" than their neighbors and hence draws lots of attention. Following the psychological finding, many approaches have focused on the detection of feature contrasts to trigger human vision nerves [19]. This research field is usually called visual attention detection or salient object detection. Liu, *et al.* [20] 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. When the saliency concept is moved from static image domain to video sequences, the motion saliency, defined as attention due to motion [21], will dominate the frames 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. Mahadevan and Vasconcelos proposed an algorithm for spatio-temporal saliency based on a center-surround framework [15]. The algorithm combined spatial and temporal components of saliency in a principled manner, and is completely unsupervised. The main shortcoming of the work was its computational performance, so it is not applicable in real time. A backtrack-chain-updation split algorithm was proposed in [22] 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.* [23] extended Chen's work to process more generalized overlapped situations. In [24], a spatio-temporal video attention detection technique was proposed to detect 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 center-surround coherency model to describe motion contrast computed by motion vectors obtained from the optical flow algorithm. (2) Employ the integral density algorithm to calculate the global minimum coherency as the motion region in the frame. (3) Present a multimedia retrieval framework to integrate global texture and local motion in order to enhance the existing retrieval framework that uses only global features.

The remainder of this paper is organized as follows. The motion saliency region detection framework is presented in Section II. Section III describes the new semantic retrieval model that fuses the global texture and local motion features to enhance the retrieval performance. The new content-based multimedia retrieval framework is also introduced in this section. Section IV presents the experimental results and analyzes the performance on KTH and TRECVID 2010 data sets from the detection and retrieval perspectives, respectively. Section V concludes the proposed motion saliency detection and semantic retrieval model.

II. Motion Saliency Region Detection

The studies on the human vision system reveal that it perceives external features separately and is sensitive to the diversity of the target region and its neighborhood [25][26]. The center-surround mechanisms of biological systems support the idea of motion saliency detection on the measurements of local motion contrast. In order to build an unsupervised detection framework on motion saliency while avoiding the "global background model" or any type of training processing, a center-surround coherency model is proposed in our proposed framework (as shown in Fig. 2)

to measure the motion contrast of a local region and its neighborhoods. After that, the integral density algorithm is utilized to achieve global minimum coherency as the motion region.

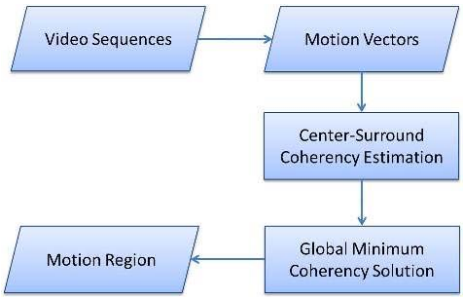


Figure 2. Motion region detection model

It is not necessary to train samples or pre-build a "global background model" for the testing instances in the proposed model. Instead, local motion information can be utilized to compute the motion saliency, so that the model could immediately adapt to different kinds of unknown backgrounds. Moreover, the model is robust to the camera motion and dynamic background because of the exploration of the global minimum coherency.

A. Motion Vector by Optical Flow

The concept of optical flows was introduced by James J. Gibson in the 1940s to describe the visual stimulus provided to animals moving through the world. In 1981, Horn and Schunck [27][28] conducted a performance analysis of a number of optical flow techniques. Recently the term optical flow has been co-opted to incorporate related techniques from image processing and control of navigation, such as motion detection, object segmentation, etc. The optical flow methods try to calculate the motion between two image frames which are taken at times t and $t + \Delta t$. If the gray value of a pixel on the image with its coordinates at time t is $I(x, y, t)$ and the pixel moves to new position at time $(t+\Delta t)$, its location on the image becomes $(x+\Delta x, y+\Delta y)$, and the gray value becomes $I(x + \Delta x, y + \Delta y, t + \Delta t)$. Assuming that the intensity is conserved, we can have Eq. (1) which can be re-written as Eq. (2). The gradient constraint equation is easily derived from a Tayor expansion of Eq. (2) as shown in Eq. (3).

$$dI(x, y, t)/dt = 0; \tag{1}$$

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t); \tag{2}$$

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t} = 0. \tag{3}$$

Let the two components of the optical flow along the x and y coordinates be $u = dx/dt$ and $v = dy/dt$, and let I_x, I_y, and I_t denote the partial x coordinate, partial y coordinate, and partial time derivatives of $I(x, y, t)$. Eq. (4) presents the basic optical flow equation.

$$I_x u + I_y v + I_t = 0. \tag{4}$$

The optical flow does not require any a priori knowledge on the object appearance, which is an important merit. However, its complex computation time makes it unsuitable for real-time applications (if without special hardware). To address such an issue, in this paper, motion vectors are used to decrease its processing time, i.e., to use the optical flow technique on the block-level motion instead of the pixel-level one. An integral part of many video compression algorithms is the motion vectors since they are used for motion compensation. The idea of the so-called block matching is to divide the current frame into a matrix of blocks that are then compared with the corresponding block and its neighbors in the previous frame to determine the motion vector. In other words, the motion vector is calculated using the optical flow method, but the motion information of the frame is presented in the block-level.

(a) (b)

Figure 3. Illustration of a center-surround sample window. (a) the original frame; (b) an illustrated center-surround region. The red area denotes the center region, and the yellow area denotes the surround region.

B. Center-Surround Coherency Model

To deal with the issues raised by the camera motion or the dynamic background, a center-surround coherency model is presented, which enables the automatic adaption of the background variations. There is no need to build or train a global model of the background. Because coherency compares the center and surrounded regions, it depends only on the relative disparity between the motion values, and therefore the new model is invariant to the camera motion.

Suppose an image is divided into $m \times n$ blocks, $1 \leq i \leq m$, $1 \leq j \leq n$, so $u_{i,j}$ and $v_{i,j}$ denote two components of the motion vectors at block (i, j). Given a center-surround region R which includes a center region R_c and a surrounded region R_s as shown in the red and yellow areas in Fig. 3(b). If the block $(i, j) \subset R$, then block (i, j) belongs to either R_c or R_s, where $1 < i_1 \leq i \leq i_4 < m$, $1 < j_1 \leq j \leq j_4 < n$. Here, i_1, j_1, i_4, and j_4 are the block boundary coordinates of R; while i_2, j_2, i_3, and j_3 are the block boundary coordinates

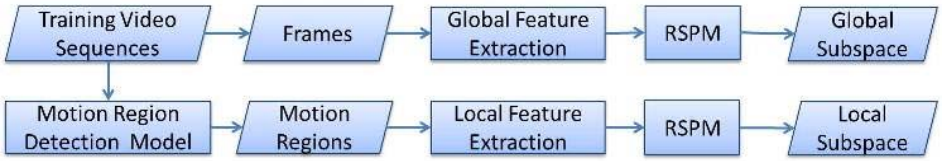


Figure 4. Training phase in the proposed semantic retrieval framework

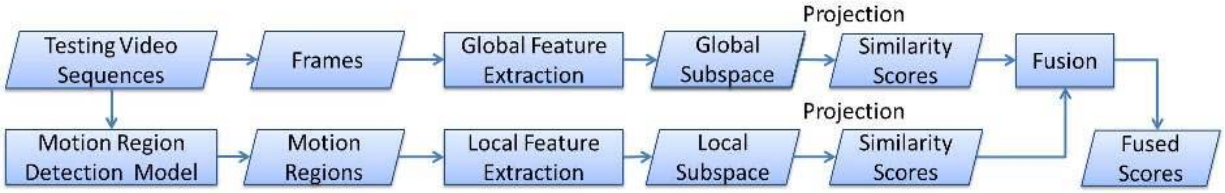


Figure 5. Testing phase in the proposed semantic retrieval framework

of R_c. The motion vectors U_c and V_c of the center region R_c are computed by summing up the motion vectors of the blocks located in the center region as shown in Eq. (5), where $\forall$ block $(i,j) \subset R_c$.

$$\begin{cases} U_c = \sum u_{i,j}; \\ V_c = \sum v_{i,j}. \end{cases} \quad i_2 \le i \le i_3;\ j_2 \le j \le j_3. \tag{5}$$

The motion vectors U and V of the region R are computed in the same manner by Eq. (6).

$$\begin{cases} U = \sum u_{i,j}; \\ V = \sum v_{i,j}. \end{cases} \quad i_1 \le i \le i_4;\ j_1 \le j \le j_4. \tag{6}$$

The motion vectors U_s and V_s of the surrounded region R_s are calculated by Eq. (7).

$$\begin{cases} U_s = U - U_c; \\ V_s = V - V_c; \end{cases} \tag{7}$$

The coherency C of the center region R_c and the surrounded region R_s can be obtained by computing the cosine similarity over the center and surrounded areas (as shown in Eq. (8)).

$$C = \cos\theta = \frac{M_c \cdot M_s}{\|M_c\|\|M_s\|}, \tag{8}$$

where $M_c = [U_c\ V_c]$ and $M_s = [U_s\ V_s]$ denote the motion energy values in the center region R_c and the surrounded region R_s, respectively. The smaller the value C is, the lower the probability that the center region and the surrounded region have similar motion activities.

The motion vector gives a quantitive measure of the block movement in the image. The greater cosine similarity of the two motion vectors, the more likely the two motion vectors come from the same object. Considering the temporal consistence of the object motion in continuous frame sequences, the sum of the coherency C_t in $\triangle t$ time window is calculated as the estimation criterion of the motion activity in region R in Eq. (9).

$$C_t = \sum_t^{t+\triangle t} C. \tag{9}$$

C. Global Minimum Coherency

After the discussion of the temporal coherency C_t, how to quickly find the global minimum coherency region in the video frame turns into an urgent problem in the unsupervised motion detection topic. Such a problem is a global search issue, which is usually very time-consuming. To solve this issue, a quick search method is presented to find the possible motion regions that have a low center-surround coherency. The integral density concept in [29], which was developed based on the integral images in [30], is adopted for it allowing a fast implementation of the box type convolution filters. Each entry in the summed area table $I_{\sum(\mathbf{x})}$ at a location **x**=(x,y) represents the sum of all the values in the input $2D$ matrix I of a rectangular region formed by the point **x** and the origin (please see Eq. (10)). After $I_{\sum(\mathbf{x})}$ is calculated, the calculation of the sum of the values over any upright rectangular areas, independent of their sizes, will take only four additions.

$$I_{\sum(\mathbf{x})} = \sum_{i=0}^{i\le x}\sum_{j=0}^{j\le y} I(i,j). \tag{10}$$

Inspired by the summed area table algorithm, the motion vectors of each block in an image are written as a matrix. The summed area table of this motion matrix is then generated for the fast computation of the center-surround coherency of every location. After traversing the center-surrounded region in the frame, the global minimum coherency can be quickly obtained.

III. Semantic retrieval model

Based on the proposed motion saliency region detection algorithm, a semantic retrieval model is presented. It consists of a new multimedia semantic retrieval framework that integrates the global texture and local motion features to enhance the retrieval performance. The motivation of this framework is to utilize the information obtained from the motion saliency region 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. 4, the training phase of the retrieval framework includes two main modules: feature extraction and subspace training, which work on the motion regions and original frames, respectively. The representative subspace projection modeling ($RSPM$) algorithm [31] is adopted to train the subspace in this proposed multimedia semantic retrieval framework. That is, a subspace called the local subspace will be trained for the local features extracted from the motion regions, and a subspace called global subspace will be trained for the global features extracted from the original video frames.

In the testing phase given in Fig. 5, 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 motion regions in the testing data set will be projected onto the local subspace obtained in the training phase (from the motion 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 the local and global subspaces to give a final similarity score to represent each video shot. A good fusion strategy can further improve the final performance of the semantic retrieval framework. In this paper, the logistic regression algorithm is employed to combine the global and local similarity scores. In future, more fusion methods will be explored in our proposed model.

IV. Experimental results and analyses

We use two data sets, KTH [32] and TRECVID 2010 (in semantic indexing task) [33], to evaluate the performance of the proposed framework. In the KTH data set, there are 25 actors performing six actions four times in four different environments with a total number of 599 video sequences. There are six action categories, namely boxing, hand clapping, hand waving, jogging, walking, and running. One characteristic of these video sequences is that they were recorded in a controlled setting with slight camera motion and a simple background.

Table I
TRECVID 2010 Data Set Used in the Experiments

Concept ID	Concept name	Number of shots
4	Airplane-flying	196
6	Animal	1816
13	Bicycling	175
38	Dancing	666
59	Hand	1053
100	Running	890
111	Sports	1299
127	Walking	3087
	Total	9182

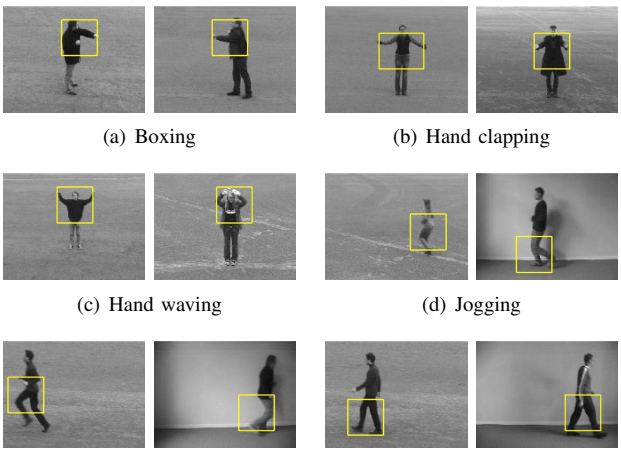
On the other hand, the data set in the semantic indexing task of TRECVID 2010 contains 130 queries, while the majority belongs to static concepts. Eight queries describing moving objects were chosen to build a subset for testing our framework, namely airplane flying, animal, bicycling, dancing, hand, running, sports, walking. These all involve salient motion. More detailed information is shown in Table I.

A. Experiments on the KTH data set

This KTH data set is used to demonstrate that the proposed framework is able to achieve pretty good performance in videos recorded in a "clean" background, even though the proposed framework is designed to deal with videos captured in uncontrolled environments. In the frame extraction step, we did not use a keyframe extraction algorithm to select the representative keyframes such as in [34]. Instead, three frames per second in average are used to compute the motion saliency in the KTH data set.

First, the accurate localization of an action is verified. Samples of motion saliency regions are illustrated in yellow boxes in Fig. 6. We notice that the motion saliency of the human body is accurately identified from the videos, while the static part of the body are excluded from the boxes. This property of the motion saliency detection model will later be transferred to the advantage of moving object-level feature extraction, and proved helpful for semantic retrieval.

In the experiments, we test the precision of the concept retrieval using those features extracted in frame-wide and region-wide, respectively. To avoid the feature bias, three kinds of texture features (Gabor, LBP, and HOG) are employed to represent each frame and motion 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 the summarization of the binary numbers, 59 LBP features are returned to represent the frame or region. Histogram of Oriented Gradients (HOG) are feature descriptors and used in computer vision and image processing. HOG features count the occurrences of



(a) Boxing (b) Hand clapping

(c) Hand waving (d) Jogging

(e) Running (f) Walking

Figure 6. Samples of motion saliency detection on KTH data set

the gradient orientation in the localized portions of an image. It is computed on a dense grid of uniformly spaced cells and uses the overlapping local contrast normalization for improved accuracy. The dimension of the HOG features used in the experiment is 135.

The Mean Average Precision (MAP) value is used to evaluate the performance of different approaches in the paper. MAP is the mean of the Average Precision (AP) of all queries. For approaches that return a ranked sequence of video shots, the AP value is a criterion that considers the order in which the returned shots are presented. In the other word, AP is the precision value averaged across all recall values between 0 and 1. Let k be the rank in the sequence of retrieved shots, n be the number of retrieved shots, $P(k)$ be the precision at cut-off k in the list, and $rel(k)$ is an indicator function with 1 if the item at rank k is a relevant shot, and 0 otherwise [35]. AP is defined as

$$AP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{number\ of\ relevant\ shots} \tag{11}$$

In the KTH data set, we select the frames of 'person' 01 - 15 as the training set and the frames of 'person' 16 - 25 as the testing set. In each frame, Gabor, HOG, and LBP features are extracted in frame-wide and region-wide, respectively. The purpose of extracting frame-wide features is to estimate the performance of only using global features and ignoring the object-level features. Then, the features of the object-level are estimated as region-wide features. We expect the features from a small area, but the motion saliency should be more discriminative than those from frame-wide.

Meanwhile, considering the possibility of the complementary information among different methods, we also test the performance of the fused similarity scores of frame-wide and region-wide. The scores are fused by Logistic Regression (LR) method. Tables III, IV, and II present the retrieval

Table II
MAP COMPARISON WHEN DIFFERENT NUMBERS OF VIDEOS ARE RETRIEVED (%) - LBP FEATURES

LBP	5	10	20	50	100
Frame-wide	3.33	5.42	9.42	15.09	18.14
Region-wide	32.36	39.48	40.82	43.29	44.24
Fused	49.17	49.51	47.29	42.24	41.08
Impr. % to frame-wide	1376.58	813.47	402.02	179.92	126.46
Impr. % to region-wide	51.95	25.41	15.85	−2.43	−7.14

Table III
MAP COMPARISON WHEN DIFFERENT NUMBERS OF VIDEOS ARE RETRIEVED (%) - GABOR FEATURES

Gabor	5	10	20	50	100
Frame-wide	52.99	53.88	48.10	43.02	39.02
Region-wide	42.92	36.60	33.88	33.13	33.56
Fused	67.78	65.07	54.34	52.01	47.25
Impr. % to frame-wide	27.91	20.77	12.97	20.90	21.09
Impr. % to region-wide	57.92	77.79	60.39	56.99	40.79

results in terms of MAP. The columns show the number of videos requested in each method. Note that the region-wide method outperforms the frame-wide one using the LBP features, while using Gabor features, the frame-wide method exceeds the region-wide one. For HOG features, if retrieving the top 5, 10, or 20 related videos, the region-wide method performs better than frame-wide one; while if retrieving more than 50 related videos, the frame-wide approach obtains a higher MAP. This result indicates that a single method does not achieve good precision on all kinds of features. Thus, a fusion technique is utilized to integrates the advantages of frame-wide and region-wide methods.

The experimental results of the fused method (labeled as "Fused") are shown in Tables IV, III, and II. The last two rows of Tables IV, III, and II list the improvement of the fused method compared to the frame-wide and region-wide methods, respectively. It can be observed that the fused method prominently improves the retrieval performance.

The average improvements of the fused method by using the Gabor and HOG features are 39.75% and 14.52%, respectively. For the LBP features, the poor performances of the frame-wide method affect the fusion results, resulting in the decrease in MAP comparing to the region-wide method in the top 50 and 100 retrieved videos. However, in the top 5, 10, and 20, the fused method achieves an increase in MAP though the performances of the frame-wide and region-wide methods are not commensurable. The overall performance of the fused method verifies that the global (frame-wide) and local (region-wide) information has the complementary discriminative potential for information retrieval.

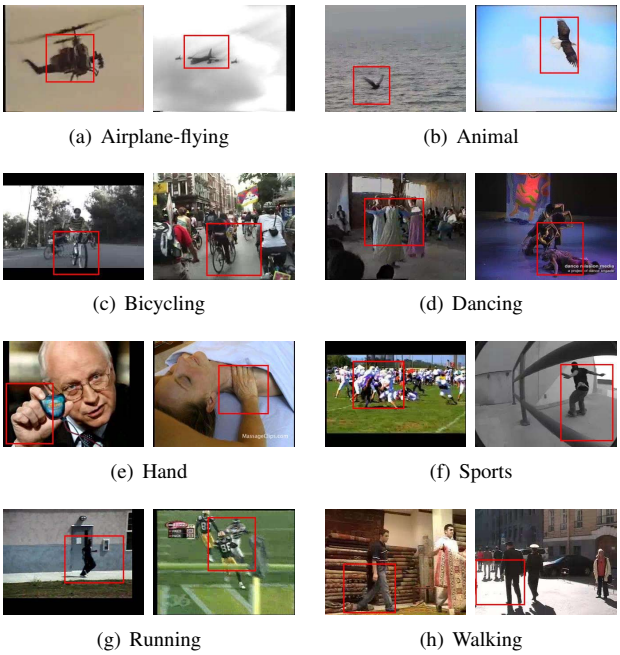
B. Experiments on the TRECVID 2010 data set

TRECVID 2010 video collection provides one reference key frame (RKF) per shot to represent the content of the

Table IV
MAP COMPARISON WHEN DIFFERENT NUMBERS OF VIDEOS ARE RETRIEVED (%) - HOG FEATURES

HOG	5	10	20	50	100
Frame-wide	57.57	62.58	67.07	68.06	65.25
Region-wide	70.09	67.35	67.33	63.91	62.10
Fused	78.68	78.14	74.33	71.54	69.09
Impr. % to frame-wide	36.67	24.86	10.82	5.11	5.89
Impr. % to region-wide	12.26	16.02	10.40	11.94	11.26

video shot. The proposed framework first computes the optical flow field on the RKF and the estimate the motion region using the center-surround coherency model. For the fast computation purpose, the searching pace in the integral density method is set to 0.05 times of the shorter dimension size of the frame and the minimum side-length of the motion region 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 target.



(a) Airplane-flying (b) Animal

(c) Bicycling (d) Dancing

(e) Hand (f) Sports

(g) Running (h) Walking

Figure 7. Some results of motion saliency region detection

Fig. 7 shows several detection results of motion saliency. The RKFs in the first and third columns come from TRECVID 2010 training data set; while the ones in the second and fourth columns are from TRECVID 2010 testing data set. These RKFs are extracted from videos containing non-static background, and most of them have camera motion and background clutter. Please note that as an unsupervised motion region detection framework, the proposed motion saliency region detection algorithm successfully identifies the main motion region in various backgrounds. This provides a good foundation for the further semantic retrieval task which views the motion regions as a kind of local information that describes the object-level texture of the shot. This may be complementary to the global information for multimedia semantic retrieval task.

V. CONCLUSIONS

Inspired by the biological mechanisms of human visions that motion saliency attracts more attention than other low-level visual features in videos, a new semantic retrieval framework for videos in non-static background is proposed, based on a novel motion saliency region detection algorithm. This framework defines a center-surround coherency model to describe the motion contrast computed by the motion vectors obtained via the optical flow algorithm, and it utilizes the integral density algorithm to calculate the global optical minimum coherency as the motion region in the frame. Further, our semantic retrieval framework integrates the global texture and local motion information obtained from the proposed motion region detection method in order to enhance the existing retrieval framework that uses only the global features.

REFERENCES

[1] S.-C. Chen, M.-L. Shyu, C. Zhang, and M. Chen, "A multimodal data mining framework for soccer goal detection based on decision tree logic," *International Journal of Computer Applications in Technology*, vol. 27, no. 4, pp. 312–323, 2006.

[2] M.-L. Shyu, C. Haruechaiyasak, and S.-C. Chen, "Category cluster discovery from distributed www directories," *Information Sciences*, vol. 155, no. 3, pp. 181–197, 2003.

[3] H. Ha, S.-C. Chen, Y. Zhu, S. Luis, S. Graham, and S. Vassigh, "Constraint driven model using correlation and collaborative filtering for sustainable building," in *Proc. of the IEEE Information Reuse and Integration (IRI)*, August 2012, pp. 309–315.

[4] M.-L. Shyu, S.-C. Chen, M. Chen, and C. Zhang, "A unified framework for image database clustering and content-based retrieval," in *Proc. of the 2nd ACM international Workshop on Multimedia databases*, no. 13, 2004, pp. 19–27.

[5] T. Meng and M.-L. Shyu, "Leveraging concept association network for multimedia rare concept mining and retrieval," in *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, Melbourne, Australia, July 2012, pp. 860–865.

[6] X. Wu, C. Zhang, and Q. Peng, "Study on feature trajectory for web video event mining," in *Rough Sets and Current Trends in Computing*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7413, pp. 219–228.

[7] C. Zhang, X. Wu, M.-L. Shyu, and Q. Peng, "A novel web video event mining framework with the integration of correlation and co-occurrence information," *Journal of Computer Science and Technology (JCST)*, in press.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] Y. Yang, H.-Y. Ha, F. Fleites, S.-C. Chen, and S. Luis, "Hierarchical disaster image classification for situation report enhancement," in *Proc. of the IEEE Information Reuse and Integration (IRI)*, August 2011, pp. 181–186.

[14] Q. Zhu, L. Lin, M.-L. Shyu, and D. Liu, "Utilizing context information to enhance content-based image classification," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 2, no. 3, pp. 34–51, 2011.

[15] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 1, pp. 171–177, 2010.

[16] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proc. of the 17th International Conference on Pattern Recognition*, vol. 2, 2004, pp. 28–31.

[17] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 10, pp. 1337–1342, 2003.

[18] R. Born, J. M. Groh, R. Zhao, and S. J. Lukasewycz, "Segregation of object and background motion in visual area mt: Effects of microstimulation on eye movements," *Neuron*, vol. 26, pp. 725–734, 2000.

[19] D. Liu, M.-L. Shyu, and G. Zhao, "Spatial-temporal motion information integration for action detection and recognition in non-static background," in *Proc. of the IEEE Information Reuse and Integration (IRI)*, August 2013, in press.

[20] 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.

[21] D. Mahapatra, S. Winkler, and S. Yen, "Motion saliency outweighs other low-level features while watching videos," in *Proc. of SPIE*, vol. 6806, 2008, pp. 68 060P1–68 060P10.

[22] 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.

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

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

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

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

[27] J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994.

[28] B. Horn and B. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1, pp. 185–203, 1981.

[29] D. Liu and M.-L. Shyu, "Effective moving object detection and retrieval via integrating spatial-temporal multimedia information," in *Proc. of the IEEE International Symposium on Multimedia (ISM)*, 2012, pp. 364–371.

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

[31] 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.

[32] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Proc. of the 17th International Conference on Pattern Recognition*, vol. 3, 2004, pp. 32–36.

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

[34] D. Liu, M.-L. Shyu, C. Chen, and S.-C. Chen, "Integration of global and local information in videos for key frame extraction," in *Proc. of the IEEE Information Reuse and Integration (IRI)*, 2010, pp. 171–176.

[35] M. Zhu, "Recall, precision and average precision," *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, 2004.