Dynamic region selection in video based on spatio-temporal multiple instance learning

XiaoZheng Wang, XuDong Zhao, Peng Liu and XiangLong Tang
School of Computer Science and Technology
Harbin Institute of Technology, Harbin 150001 China
wangxiaozheng@foxmail.com, {zhaoxudong, pengliu, tangxl}@hit.edu.cn

Abstract

The selection of dynamic region in video plays an important role in many subsequent vision-based applications, especially in scene classification with different weather conditions. In this paper, we extract five local features from pixel blocks of each frame in a video, and propose an approach to dynamic region selection based on a presented description of spatio-temporal multiple instances. The effectiveness of our method is shown using experiments on videos under different weather environments.

Keywords: *Multi-instance, spatio-temporal feature, dynamic region, K-means.*

1. INTRODUCTION

Dynamic region selection aims at finding the significant regions that are composed of the locations containing obvious changes in video. It plays an important role in many applications (e.g., motion coding [2], motion detection [5], scene modeling [4], scene classification [3], weather classification [6] and etc.). In the same way, Multiple Instance Learning (MIL) [1] concerns labels of instances included in each bag to classify bags. In video, bags correspond to image patches over a time slot; while, instances refer to spatial or temporal features.

In this paper, we propose a dynamic region selection approach based on spatio-temporal multiple instance learning using five local features from a pixel block. First of all, we subdivide a frame image into 10×10 pixel blocks, and extract features to form spatial multiple instances. Then, we aggregate features frame by frame in each pixel block to produce temporal multiple instances. Together, the spatio-temporal multiple instances consist of a bag that corresponds to a pixel block. Finally, K-means clustering of bags is used to select dynamic region in video. The organization structure is illustrated in Fig. 1.

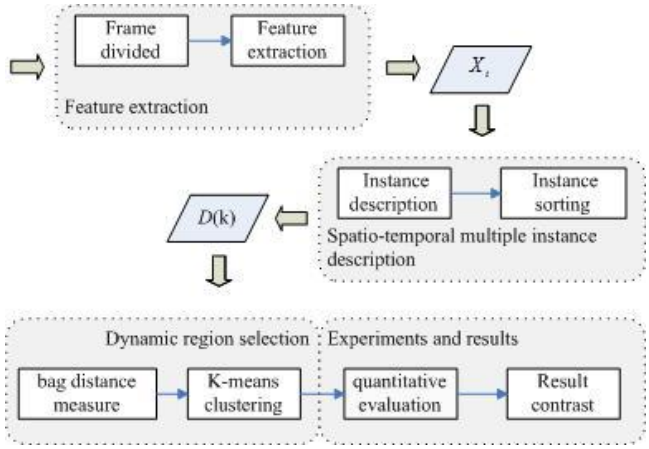

Fig.1. Framework of spatio-temporal multiple instance based dynamic region selection in video

2. FEATURE EXTRACTION

Features are derived from a pixel block. First, we divide an image into 10×10 pixel blocks. Five features are extracted in each block, i.e., hue, saturation, minimum brightness, local contrast and sharpness. Gray-scale features are commonly used for image processing tasks that range from low level algorithms to highly sophisticated modules. However, we pay more attention to color information according to the low visibility deriving from bad weather. Compared to RGB color space, HSV space keeps the same way on perception of color information that human eye does. So we extract hue, saturation and brightness at pixels. The minimum value of brightness ($V_{\min}$) and the local mean values of hue (H) and saturation (S) are taken in each pixel block. To increase the robustness of contrast estimation, we define the local contrast as follows,

$$C = \frac{V_{\max} - V_{\min}}{V_{\max} + V_{\min}} \quad , \tag{1}$$

where $C, V_{\min}$ and $V_{\max}$ represent the local contrast, the minimum and the maximum value of brightness, respectively. Besides, clearly distinguishable objects under fine weather conditions are expected to have sharp edges with large contrast differences. In addition to the contrast feature discussed above, a gradient-based method is used to determine the sharpness of the test images. It is based on an average determination of the sobel gradient magnitude, which is defined as follows,

$$T = \frac{\sum_i \sqrt{S_X^2(i) + S_Y^2(i)}}{\sum_i 1} \quad . \tag{2}$$

The sharpness T is derived from an average determination of the sobel gradient magnitude S_X and S_Y with i belonging to a pixel block.

3. MULTIPLE INSTANCE DESCRIPTION

Multiple-instance learning (MIL) is a variation on supervised learning. Instead of receiving a set of instances which are labeled positive or negative, the learner receives a set of bags that are labeled positive or negative. Each bag contains many instances. The most common assumption is that a bag is labeled negative if all the instances in it are negative. On the other hand, a bag is labeled positive if there is at least one instance in it which is positive. From a collection of labeled bags, the learner tries to either induce a concept that will label individual instances correctly or learn how to label bags without inducing the concept. For a more accurate and detailed expression of spatial features in each block, we subdivide each 10×10 pixel block into four 5×5 pixel blocks in space, and extract features from every 5×5 block. In other words, we get spatial feature vectors X_t from a 10×10 block. Let X_t be $X_t = \{x_t^1, x_t^2, x_t^3, x_t^4\}$. t represents the current frame.

The feature vector from a 5×5 pixel sub-block x_t^m $(m=1,2,3,4)$ is expressed as $x_t^m=(H,S,V_{\min},C,T)$. Furthermore, we aggregate feature vectors frame by frame. Assuming the frame number to be *M*, each 10×10 block contains *M* spatial multiple instances, which is recorded as $X_1, X_2, \ldots X_t \ldots, X_M$. In order to observe the changes in each block more clearly, we sort these M temporal multiple instances from largest to smallest by calculating the modulus of the obtained spatial vectors, which is expressed as $|X_t|=\frac{1}{4}\left(|x_t^1|+|x_t^2|+|x_t^3|+|x_t^4|\right)$. Thus, a new sequence of temporal instances in a bag is expressed as $Y_1, Y_2, \ldots, Y_M$, where Y_t is $Y_t=\{y_t^1, y_t^2, y_t^3, y_t^4\}$. In addition, a corresponding subtraction operation to different temporal instances in each bag is made, i.e., $d_i^m=y_i^m-y_j^m$ $(i\in[1,M], j\in[i+1,M])$. Let y_i^m be $y_i^m\in Y_i$ and y_j^m be $y_j^m\in Y_j$. Besides, we sort these spatio-temporal vectors from largest to smallest by calculating the modulus $|d_i^m|$ to get a sequence $b_1, b_2, ..., b_N, ..., b_{2M(M-1)}$. In order to reduce the complexity of the algorithm, we select the first *N* (*N*=*M*/4) vectors as the spatio-temporal multiple instances. The bag of a pixel block is expressed as $D(k)=\{b_1(k), b_2(k), ..., b_N(k)\}$, where $k\in[1,n]$. *n* denotes the number of 10×10 pixel blocks in an image.

4. UNSUPERVISED CLASSIFICATION

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume *k* clusters) fixed a priori. The main idea is to define *k* centroids, one for each cluster. In this paper, we define *k* as 2. These two centroids shoud be placed in a cunning way because different location causes different result. So, the better choice is to place them as much as possible far away from each other. *K*-means clustering is utilized to accomplish spatio-temporal multiple instance learning. Bags representing locations containing obvious changes in video are selected as dynamic region. Two different distance metrics between bags based on Hausdorff distance (HD) are proposed. The max HD for classification of bags in video is expressed as follows,

$$\begin{aligned} maxh(D(k),D(r)) &= \max_{b_p(k)\in D(k)} \min_{b_q(r)\in D(r)} \left\| b_p(k)-b_q(r) \right\| \\ maxh(D(r),D(k)) &= \max_{b_q(r)\in D(r)} \min_{b_p(k)\in D(k)} \left\| b_p(k)-b_q(r) \right\| \end{aligned}, \quad (3)$$

where $k\in[1,n]$ and $p,q\in[1,\mathrm{N}]$. $\|\|$ denotes a norm. The HD expressed in Equation (3) is a directed distance. Therefore, the max HD distance between different bags is

$$maxH(D(k),D(r))=\max\{maxh(D(k),D(r)), maxh(D(r),D(k))\}. \quad (4)$$

The average HD distance to classify different bags is

$$\begin{aligned} avgh(D(k),D(r)) &= \frac{1}{|D(k)|}\sum_{b_p(k)\in D(k)} \min_{b_q(r)\in D(r)} \left\| b_p(k)-b_q(r) \right\| \\ avgh(D(r),D(k)) &= \frac{1}{|D(r)|}\sum_{b_q(r)\in D(r)} \min_{b_p(k)\in D(k)} \left\| b_p(k)-b_q(r) \right\| \end{aligned}. \quad (5)$$

avgh is also a directed distance. Thus, we define a maximal average of HD as follows,

$$maxavgH(D(k),D(r))=\max\{avgh(D(k),D(r)), avgh(D(r),D(k))\}. \quad (6)$$

Then, we follow the follow *K*-means clustering step as shown in Algo.1 to accomplish a classification of bags representing the static and dynamic region in video.

Input: We define the sorted collections of bags as *U*, and the number of the clusters *K* as 2.

Output: The label of each bag and the centroid of every cluster.

1. Select *K* bags from collections *U* randomly as the initial centroids $C_1, C_2, \ldots, C_K$. A cluster collection G_j is formed using initial centroids , where $G_j=\{C_j\}$, $j=1,2,\ldots,K$.
2. while $\forall j=1,2,\ldots,K$, we have $C_j^*\neq C_j$, do
3. $C_j=C_j^*$;
4. For $i\leftarrow1$ to n do
5. If $H(B_i,C_j)\le H(B_i,C_k)$ ($\exists j=1,2,\ldots,K, \forall k=1,2,\ldots K$, and $j\neq k$) then
6. $G_j\leftarrow B_i$; // $B_i(i=1,2,\ldots,n)$ is the data bag, G_j is the cluster collection.
7. $G_j^*=G_j\cup\{B_i\}$;
8. $G_j=G_j^*$;
9. //Updating the centroid
10. $\overline{B_j}=\frac{1}{|G_j|}\sum_{B_q\in G_j} B_q$;
11. End if
12. End for
13. End while

5. EXPERIMENTS AND RESULTS

We have tested our method on two video clips with obvious dynamic regions. One is a privately shot video clip with fast lighting change (namely, DRS_FLC). The other is a public movie clip with snow (namely, DRS_S). Temporal multiple instance learning method based on intensity differences is utilized for qualitative comparison. Moreover, the ground truth of dynamic regions in each video is manually labeled for quantitative analysis. *TP* and *FP* represent the truly and falsely segmented dynamic region. Meanwhile, *TN* and *FN* denote the true and false static region. Furthermore, *Precision* and *Recall* are respectively defined as *Precision=TP/(TP+FP)* and *Recall=TP/(TP+FN)*. The experimental results are shown in Fig. 2, Fig. 3 and Table 1. It can be observed that max avgH is more competent for dynamic region selection in video with snow than maxH. On the contrary, maxH works better on classification of dynamic and static bags in video with fast lighting change than max avgH.

6. CONCLUSION

In this paper, a dynamic region selection approach in video is proposed based on spatio-temporal multiple instance learning. Considering the influence of different weather environments, we firstly extract spatial feature vectors from sub-blocks in a pixel block. After that, temporal aggregation of spatial feature vectors is performed for the constitution of temporal multiple instances. By means of the subtraction operation, we obtain an instance bag which could present the pixel change in each block. Finally, *K*-means clustering of these bags is used to select dynamic areas

in video based on two different Hausdorff distance measures. Experiments indicate the effectiveness of our method.

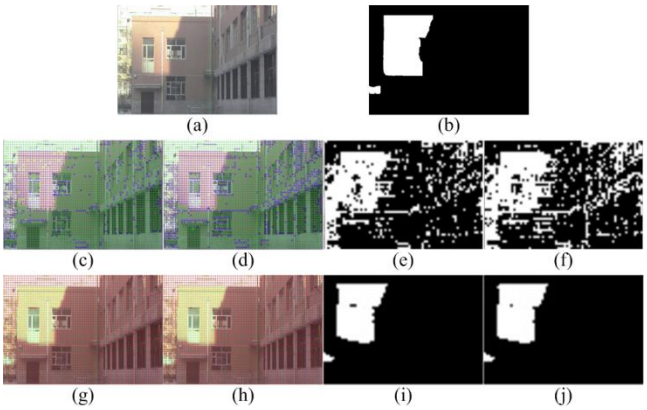

Fig. 2 Dynamic region selection on DRS_FLC
(a) Video clip with fast light change
(b) Labeled ground truth
(c) Result of MIL based on temporal intensity differences using maxH
(d) Result of MIL based on temporal intensity differences using max avgH
(e) Corresponding binary result using maxH
(f) Corresponding binary result using max avgH
(g) Result of spatio-temporal MIL using maxH
(h) Result of spatio-temporal MIL using max avgH
(i) Corresponding binary result using maxH
(j) Corresponding binary result using max avgH

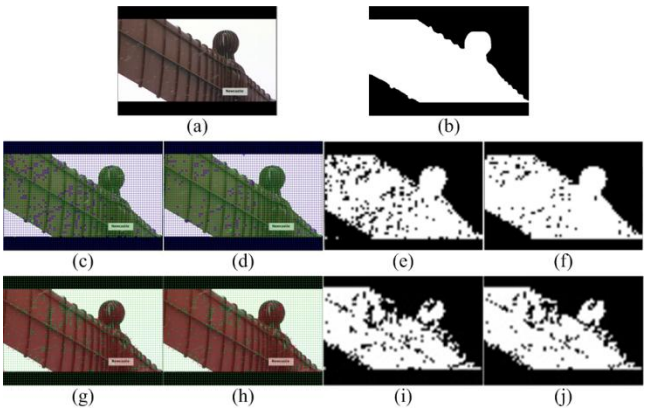

Fig. 3 Dynamic region selection on DRS_S
(a) Video clip with snow
(b) Labeled ground truth
(c) Result of MIL based on temporal intensity differences using maxH
(d) Result of MIL based on temporal intensity differences using max avgH
(e) Corresponding binary result using maxH
(f) Corresponding binary result using max avgH
(g) Result of spatio-temporal MIL using maxH
(h) Result of spatio-temporal MIL using max avgH
(i) Corresponding binary result using maxH
(j) Corresponding binary result using max avgH

7. REFERENCES

[1] Ali, S., Shah, M.: 'Human action recognition in videos using kinematic features and multiple instance learning', *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2010, 32, (2), pp. 288-303

[2] Ascenso, J., Brites, C., Pereira, F.: 'Content adaptive Wyner-Ziv video coding driven by motion activity', *IEEE International Conference on Image Processing (ICIP)*, October 2006, Atlanta, GA, pp. 605-608 AuthorC. Reference3.

[3] Bosch, A., Zisserman, A., Muoz, X.: 'Scene classification using a hybrid generative/discriminative approach', *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2008, 30, (4), pp. 712-727

[4] Grech, R., Monekosso, D., Remagnino, P.: 'Building visual memories of video streams', *Electronics Letters*, 2012, 48, (9), pp. 487-488

[5] Kim, T., Im, J., Paik, J.: 'Video object segmentation and its salient motion detection using adaptive background generation', *Electronics Letters*, 2009, 45, (11), pp. 542-543

[6] Roser, M., Moosmann, F.: 'Classification of weather situations on single color images', *IEEE Intelligent Vehicles Symposium*, June 2008, Eindhoven, Netherlands, pp. 480-485

		TP	FP	TN	FN	Precision	Recall
DRS _S	maxH	873	28	1678	877	0.9689	0.4989
	maxavgH	935	26	1680	815	0.9729	0.5343
DRS _FLC	maxH	326	6	2938	186	0.9819	0.6367
	maxavgH	335	3	2941	177	0.9911	0.6543

Table 1: Comparison on a quantitative analysis

8. ABOUT THE AUTHOR

Xiaozheng Wang is pursuing M.Sc. degree in Computer Science from Harbin Institute of Technology. His current research interest lies in motion tracking and machine learning.

Xudong Zhao received his Ph.D. degree in Artificial Intelligence and Information Processing from Harbin Institute of Technology. His current research interests include statistical machine learning, pattern recognition, time series analysis and image processing.

Peng Liu received his Ph.D. degree in 2007. He is currently an Associate Professor at school of computer science and technology in Harbin Institute of Technology. His research interests include digital signal processing, pattern recognition and VLSI design.

Xianglong Tang received his PhD degree from Harbin Institute of Technology, China in 1995. He is currently a Professor at school of computer science and technology in Harbin Institute of Technology. His research interests include OCR, biometrics, image processing and pattern recognition.