Temporal multiple instance clustering for dynamic region selection in video

Yongqiang Zhang, Xudong Zhao, Daming Shi, Xianglong Tang
School of Computer Science and Technology
Harbin Institute of Technology, Harbin 150001 China
qtds0@163.com, {zhaoxudong, dshi, tangxl}@hit.edu.cn

Abstract

Video dynamic region corresponds to the selection of pixels according to their temporal value changes. Based on temporal multiple instance learning, we propose a dynamic region selection approach with three major contributions. First, a temporal bag and instance description differing from conventional multiple instance definition is made. Second, a bag distance measure is presented as an improvement for multi-instance clustering. Third, learning on clustering centers of bags is modified for rapid convergence. The effectiveness of our method is demonstrated using experiments on videos under different weather conditions.

Keywords: *Multi-instance, dynamic region selection, Hausdorff distance, K-means.*

1. INTRODUCTION

Dynamic region selection in video is utilized as a critical task for the subsequent video analysis (e.g. object tracking[1], motion coding[4], scene modeling[3] and etc.). A most common method derives from the accumulate frame differences[6], in which a manual threshold has to be appointed in advance. In fact, video region selection corresponds to the problem of Multiple Instance Learning (MIL, for short)[2]. MIL concerns the labels of the instances included in each bag to classify bags. As to video processing, regions and included image patches in a single frame are considered as bags and instances, respectively[1]. Ordinarily, the temporal information of a video is neglected.

In this paper, we propose a temporal description of bags and instances at pixel level based on MIL, and present a related approach without any parameters for dynamic region selection in video under different circumstances. First of all, bags and the corresponding instances are described. Then, the instances of each bag are sorted. An improved representation of the distance metric between bags is explored. A following algorithm based on *K*-means clustering is modified to accomplish dynamic region selection. Finally, experiments between the accumulated frame differences and the proposed approach are made. The organization structure is illustrated in Fig. 1.

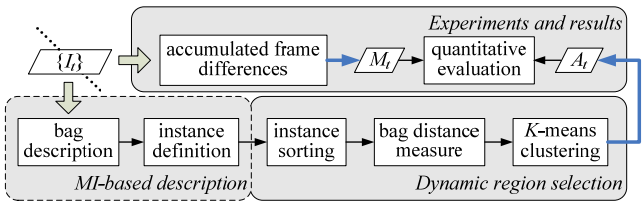


Figure 1: Framework of MIL-based dynamic region selection in video.

2. MI-BASED DESCRIPTION

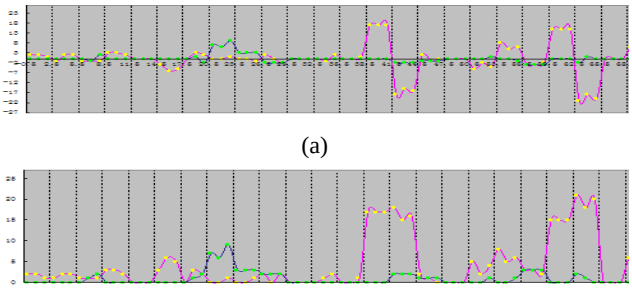
Actually, dynamic region selection consists with a generalized multi-instance problem[5], i.e., it corresponds to the classification of bags. In MIL, a positive bag possibly contains instances labeled negatively and vice versa. In outdoor video that contains different weather conditions, a location of a dynamic region might have very limited pixel changes, due to a fast lighting change existing only in a short time slot over the observation period or few snow appearances. Considering the temporal correlation of pixel values in video, we view each pixel as a bag. Therefore, dynamic region selection is equivalent to the searching of positive bags in scene. Then, a description of instance is in demand for the classification of bags. We refer to $AD_t^c(y)$ as the absolute difference of a pixel value in RGB color space ($c \in \{R,B,G\}$). That is

$$AD_t^c(y) = | I_t^c(y) - I_{t-1}^c(y) |, \quad (1)$$

where y and t represent the location and the current frame, respectively. Correspondingly, a definition of instance is expressed as follows,

$$Instance_t(y) = [AD_t^r(y), AD_t^g(y), AD_t^b(y)]. \quad (2)$$

An example of MI-based description according to two pixels (e.g., a positive pixel and a negative pixel) is shown in Fig. 2.



(a)

(b)

Figure 2: MI-based description of two pixels. (a)The original frame difference; (b)The absolute frame difference.

The three yellow points between two adjacent dotted lines represent an instance of the dynamic pixel corresponding to the difference in R, G and B channel, respectively. Similarly, the three green ones represent an instance of the static pixel.

3. DISTANCE MEASURE IN MIL

We aim at differentiating between static and dynamic bags, and meanwhile clustering bags with the same label. The fact is that each positive bag in a dynamic region probably contains static instances labeled negatively. Commonly, Hausdorff distance (HD) is selected as a standard measure between bags in MIL. Thus, the HD for classification of bags in video is expressed as follows,

$$maxh(A,B) = \max_{a\in A}\min_{b\in B}\|a-b\|$$
$$maxh(B,A) = \max_{b\in B}\min_{a\in A}\|a-b\| \quad (3)$$

where A and B represent bags. $\|\cdot\|$ denotes a norm. The HD expressed in Equation (3) is a directed distance. Therefore, we select the max HD as a distance measure. That is

$$maxH(A,B) = \max\{maxh(A,B), maxh(B,A)\}. \quad (4)$$

However, the selection of a dynamic region is conditioned by different circumstances in video. We follow our previous research [7] that the absolute difference of a pixel value in video under fast lighting change keeps a continuous intensity change. That is to say, $AD_t^{\max}(y)$ remains almost the same over the right period that corresponds to fast illumination variations. Thus, it is the *max H* expressed in Equation (4) that can be better used for real-time dynamic region selection under fast lighting change. On the contrary, *max H* is sensitive to noise (e.g., randomly distributed snow over time) because of neglecting the contribution from other instances except the farthest couple of instances from two different bags. Therefore, we additionally define an average HD to classify different bags in video. That is

$$avgh(A,B) = \frac{1}{|A|}\sum_{a\in A}\min_{b\in B}\|a-b\|$$
$$avgh(B,A) = \frac{1}{|B|}\sum_{b\in B}\min_{a\in A}\|a-b\| \quad (5)$$

where $|\cdot|$ denotes the instance number of a bag. *avgh* is also a directed distance. Thus, we define a maximal average of HD as follows,

$$maxavgH(A,B) = \max\{avgh(A,B), avgh(B,A)\}. \quad (6)$$

This distance measure takes into account all the instances, so it's robust to the noise. That is, the weight of very dynamic instances is reduced. Besides, their dynamic characteristic remains still. The difference between the maximum average HD and the maximum HD is shown in Fig. 3.

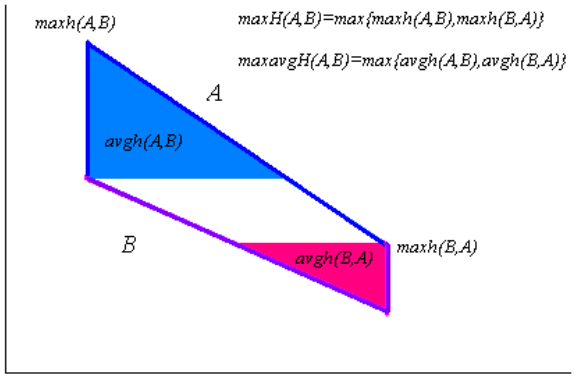


Figure 3: The difference between the maximum average HD and the maximum HD

In this figure, we select two special bags, the instances of which have been sorted correspond to the upper line *A* and lower line *B*. The maximum HD is equal to the longer edge of the two vertical lines, and the maximum average HD is equal to the larger one of the two areas.

4. SORTING AND SELECTION OF INSTANCES

The distance of each instance in bag *A* to each instance in bag *B* will be considered in order to compute HD between bag *A* and bag *B*, and vice versa. As the number of instances increases, the time cost for calculating the distance grows exponentially. In order to reduce the time cost, we sort instances of a bag in a descending order, and take the *M* first instances as a representation of the bag. Of course, we should ensure that the stability of distance measure remains after sorting. And there is another motive for sorting, i.e., we can update each center more quickly when clustering bags in the following step.

As to location y in each frame, we select its max component $AD_t^{\max}(y)$ in R, G and B channel for instance sorting. That is,

$$AD_a^{\max}(y) = \max\{AD_a^r(y), AD_a^g(y), AD_a^b(y)\}$$
$$AD_b^{\max}(y) = \max\{AD_b^r(y), AD_b^g(y), AD_b^b(y)\} \quad (7)$$

An instance a should be sorted before an instance b, if $AD_a^{\max}(y) > AD_b^{\max}(y)$. The sorting result is shown in Fig. 4.

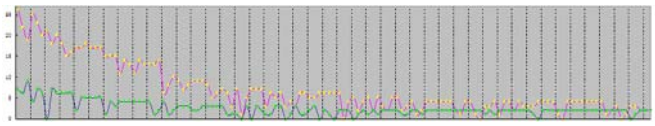
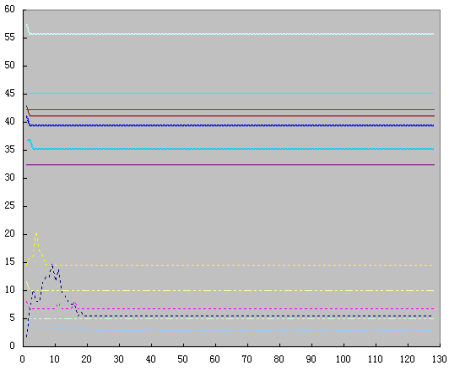
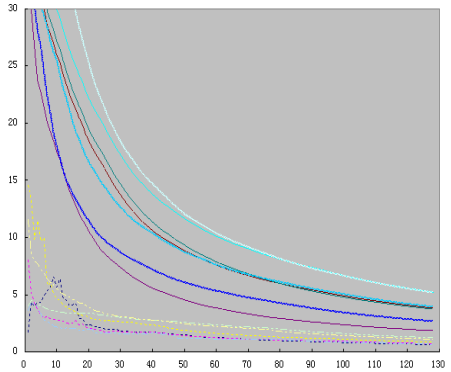


Figure 4: Sorting of instances of two bags



(a)



(b)

Figure 5: The variation of HD with the growth of the number of instances. (a) The maximum HD (b)The maximum average HD

After sorting, the maximum HD is equal to the distance from the first instance of a bag to the first instance of another bag. When the number of selected instances increases, the maximum HD keeps mostly constant.

As described in Section 3, the maximum average HD is calculated according to a directed distance *avgh*, which is the average distance of all the instances of a bag to another. Since the dynamic appearances of the low-ranking instances is less than that of the ones ranking high, the maximum average HD decreases with the growth of the number of selected instances. Yet, the relative distance between the static and dynamic bags keeps stable. Meanwhile, the influence of isolated instances is reduced.

As shown in Fig. 5, we select two groups of typical bags (a static group and a dynamic group) and calculated the maximum HD and the maximum average HD with a growing number of selected instances. In Fig. 5, the solid lines point to outer-class distances from dynamic bags to static bags. The dashed lines point to inner-class distances in static bags or in dynamic bags. Experimental result is consistent with the analysis above. With the growth of the number of selected instances, the maximum HD remains as a straight line. The maximum average HD reduces gradually as the curve $y=1/x$, but the relative position remains stable. Moreover, the outer-class distance is much larger than the inner-class distance with both of the distance measures mentioned above. It means that the maximum HD or the maximum average HD is feasible to distinguish the static bags and the dynamic ones.

5. DYNAMIC REGION SELECTION BY *K*-MEANS

We follow the prevailing *K*-means clustering step to accomplish a classification of bags representing the static and dynamic region in video. Using HD expressed in Equation (4) and Equation (6), we improve the calculation of clustering center. That is

$$\overline{B_j} = \frac{1}{|G_j|}\sum_{B_q \in G_j} B_q \quad (j = 1,2), \tag{8}$$

where B_q represents a bag in cluster G_j. $|G_j|$ denotes the bag number of G_j. That is, we consider $\overline{B_j}$ as a new clustering center.

After clustering, the pixels are divided into two groups: One group represents the dynamic pixels, the illumination variation of which is stronger. And another group corresponds to the static pixels.

6. EXPERIMENTS AND RESULTS

We have tested our method on a database named as DRS [1] containing 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). Moreover, accumulated frame differences are utilized for qualitative comparison. It is difficult to label the ground truths manually for quantitative analysis. Therefore, we set a large number of manual thresholds of accumulated frame differences, and artificially select a best dynamic region selection result as a ground truth. TP and FP represent the truly and falsely selected dynamic region. Meanwhile, TN and FN denote the true and false static region. Furthermore, we make Precision=TP/(TP+FP) and Recall=TP/(TP+FN). The experimental results are shown in Fig. 6, Fig. 7 and Table 1. A demo of the experimental results is also shown in DRS. It can be observed that *max avgH* is more competent for dynamic region selection in video with snow than *max H*. On the contrary, *max H* works better on classification of dynamic and static bags in video with fast lighting change than *max avgH*.

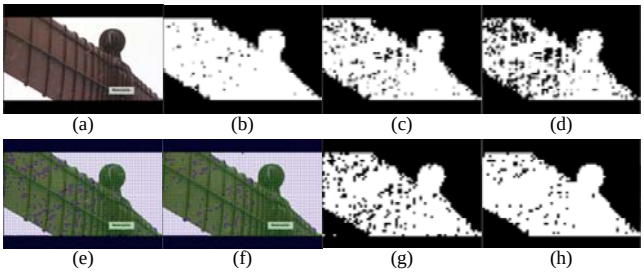


Figure 6: Dynamic region selection on DRS_S. (a) Video clip with snow; (b) Accumulated frame differences with Th=128; (c)Accumulated frame differences with Th=160 (selected); (d)Accumulated frame differences with Th=196; (e) Selection result using *maxH*; (f) Selection result using *maxavgH*; (g) Binary result using *maxH*; (h) Binary result using *maxavgH.*

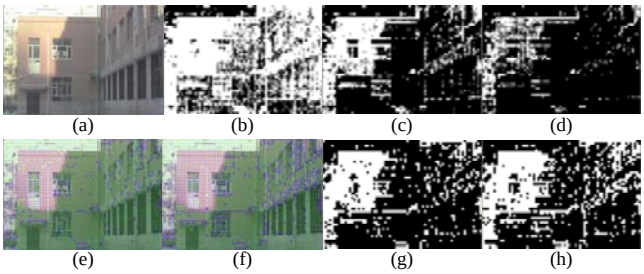


Figure 7: Dynamic region selection on DRS_FLC. (a) Video clip with fast light change; (b) Accumulated frame differences with Th=128; (c) Accumulated frame differences with Th=196 (selected); (d) Accumulated frame differences with Th=256; (e) Selection result using *maxH;* (f) Selection result using *maxavgH;* (g) Binary result using *maxH;* (h) Binary result using *maxavgH.*

Table 1: Comparison on a quantitative analysis

		TP	FP	TN	FN	Precision	Recall
DRS_S	*maxH*	1519	123	1762	52	0.925	0.967
	maxavgH	1624	18	1729	85	0.989	0.950
DRS_FLC	*maxH*	2280	173	881	122	0.929	0.949
	maxavgH	2182	271	997	26	0.890	0.988

7. CONCLUSION

In this paper, we propose a MIL-based dynamic region selection approach in video. According to pixel-wise time correlation and color information, we firstly describe bags and instances in video. After the sorting of instances in each bag, we improve a max Hausdorff distance measure and present a maximal average one adaptive to separate videos under different weather conditions. The clustering center is modified for the rapid convergence of MIL-based *K*-means clustering. Experimental results indicate the effectiveness of our method, which provides a first step support for subsequent video analysis.

[1] http://pr-ai.hit.edu.cn/percy/DRS

8. REFERENCES

[1] Babenko, B., Yang, M. H., & Belongie, S. (2011). Robust object tracking with online multiple instance learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *33*(8), 1619-1632.

[2] Dietterich, T. G., Lathrop, R. H., & Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*,89(1), 31-71.

[3] Grech, R., Monekosso, D. N., & Remagnino, P. (2012). Building visual memories of video streams. *Electronics letters*, 48(9), 487-488.

[4] Secker, A., & Taubman, D. (2004). Highly scalable video compression with scalable motion coding. *Image Processing, IEEE Transactions on*, 13(8), 1029-1041.

[5] Weidmann, N., Frank, E., & Pfahringer, B. (2003). A two-level learning method for generalized multi-instance problems. In *Machine Learning: ECML 2003* (pp. 468-479). Springer Berlin Heidelberg.

[6] Yin, H., Chai, Y., Yang, S. X., & Yang, X. (2011). Fast-moving target tracking based on mean shift and frame-difference methods. *Systems Engineering and Electronics, Journal of*, 22(4), 587-592.

[7] Zhao, X., Liu, P., Liu, J., & Tang, X. (2011, November). A time, space and color-based classification of different weather conditions. In *Visual Communications and Image Processing (VCIP), 2011 IEEE* (pp. 1-4). IEEE.

About the author

Yongqiang Zhang is pursuing M.Sc. degree in Computer Science from Harbin Institute of Technology. His current research interest lies in motion tracking.

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

Daming Shi (M'02-SM'04) received his Ph.D. degree in mechanical control from Harbin Institute of Technology, China, and the Ph.D. degree in computer science from University of Southampton, United Kingdom. He is currently a Professor at Harbin Institute of Technology, China, and served as an Assistant Professor at Nanyang Technological University, Singapore, from 2002 to 2009. His current research interests include machine learning, medical image processing, pattern recognition and neural networks.

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.