An Automatic Object Retrieval Framework for Complex Background

Yimin Yang[1], Fausto C. Fleites[1], Haohong Wang[2], and Shu-Ching Chen[1]

[1]*School of Computing and Information Sciences, Florida International University*
[3]*TCL Research America*
yyang010@cs.fiu.edu, fflei001@cs.fiu.edu, haohong.wang@tcl.com, chens@cs.fiu.edu

***Abstract*—In this paper we propose a novel framework for object retrieval based on automatic foreground object extraction and multi-layer information integration. Specifically, user interested objects are firstly detected from unconstrained videos via a multimodal cues method; then an automatic object extraction algorithm based on GrabCut is applied to separate foreground object from background. The object-level information is enhanced during the feature extraction layer by assigning different weights to foreground and background pixels respectively; and the spatial color and texture information is integrated during the similarity calculation layer. Experimental results on both benchmark data set and real-world data set demonstrate the effectiveness of the proposed framework.**

***Keywords*-object retrieval; object extraction; multi-layer information fusion;**

I. Introduction

As a conceptual level of content-based image retrieval (CBIR), object (especially from videos) retrieval has gained significant importance and attracted more and more attention. Object retrieval is not only a hot topic in academic society but also a promising practice in real-world. For example, it is not unusual that people are interested in finding the same or similar object that appears in the video they just watched. Traditional CBIR works engage in bridging the gap between low-level image features and high-level semantics by analyzing the whole content of static images without considering human interest. To put more emphasis on the potential object region, many attempts have been made to approach human perception system by segmenting images into regions and model the image content via so-called region-based local features. However, the performance is far beyond satisfactory due to the limitation of segmentation techniques and the obstacle of salient object identification especially when multiple objects are involved.

The difficulty of retrieval task escalates into another level when dealing with frames from digital videos instead of static images because videos are usually filmed under various lighting conditions in an unconstrained manner. Specifically there are three major difficulties for the task of video object retrieval. First, the potential objects of human interest in videos are accompanied by extremely noisy background with numerous variance such as deformation, occultation, rotation, scale, affine transform, and translation. Second, how to effectively and efficiently describe and represent the content in an image (video frame) is very critical for precisely retrieving the exact or similar object appeared in the video. Finally, the evaluation of an image retrieval system is relative subjective and lacks a widely acknowledged standard, which makes the improvement of object retrieval task even harder.

In this work, we have presented a novel object retrieval approach that is able to automatically extract video object from complex background and conduct efficient object retrieval by fusing spatial color and texture information. To the best of our knowledge, this is the first attempt to perform automatic video object retrieval based on the integration of concept-level spatial color and texture knowledge. The novelties of this work is summarized as follows:

- Proposes a novel multi-layer fusion method based on concept-level spatial color and texture information, where salient objects are automatically extracted from complex background for feature extraction.
- Develops a novel video object retrieval approach that seamlessly integrates automatic object extraction and semantic fusion for effective object retrieval.

The rest of the paper is organized as follows: section II introduces the related work on object segmentation and image content representation; section III discusses the details of the proposed object retrieval framework; section IV presents the experimental analysis; and section V concludes the paper.

II. Related work

There has been a lot of work on multimedia retrieval and semantic concept detection [1, 2]. Since the focus of this paper is automatic foreground object extraction and efficient content representation strategy, the related work will be introduced from these two perspectives.

Object segmentation is a fundamental problem in applications of object recognition, image classification and image/video retrieval. Many efforts have been put to this area, obtaining promising results. Carreira and Sminchisescu [3] propose an automatic object segmentation method based on constrained parametric Min-Cuts (called CPMC), which is able to automatically detect multiple objects in static natural images. Other work requiring a certain amount of manual interaction include GrabCut [4] algorithm and Bagon *et al.*'s work [5]. In this paper, we propose an automatic

object extraction approach based on the popular GrabCut algorithm without human interaction.

The work on image content representation for efficient image classification/retrieval can be roughly categorized into three classes [6]: (1) BOW-based. One major drawback of the BOW-based strategy is the neglect of spatial information. To overcome this disadvantage, much work has been done on exploring spatial context [7] and sparse coding [8]. (2) Region-based. Some representative work includes [9] using eigenregion and [10] modeling region semantics via contextual Bayesian networks. Two intrinsic problems of the region-based strategy are the limited number of object categories and the imperfection of segmentation results. (3) Fusion-based. There are three traditional fusion methodologies, i.e., early fusion (feature level), late fusion (decision level) and the combination [11]. In addition, there is another emerging kernel-based fusion method [12], which suffers from high computation cost and over fitting issue.

In this paper, we present a multi-layer fusion scheme, which jointly combines object-level early fusion and similarity-level late fusion based on the information obtained from the object extraction component, and finally constructs an automatic object retrieval framework.

III. Proposed object retrieval framework

The overall framework of the proposed object retrieval approach contains three major components, naming (1) video object extraction; (2) object-level feature extraction and similarity fusion; and (3) the final visual retrieval. In this work the first two components are the main contributions and will be elaborated in the following subsections.

A. Video Object Extraction

1) Object detection: There are existing works for detecting arbitrary object in a video given the object modal being sufficiently well trained. However, false detection may still occur, which should be taken into consideration. Fortunately, there is a refinement method for object detection in unconstrained video sequences based on multimodal cues [13]. Specifically, it combines appearance, spatial-temporal, and topological cues to aid object detection, where the appearance cue dictates the probability of object occurrence and its location in a video frame, while the spatial-temporal and topological cues reflect relational constraints between the target object class and a related object class. For example, if a bag is a target object, then the related object could be a face. The three cues are modeled respectively as follows

$$\boldsymbol{\rho}\left(O^i, O^j\right) = \begin{cases} 0 & \text{if } i = 0; \\ c\left(v\left(O^i\right), v\left(O^j\right)\right) & \text{otherwise.} \end{cases} \quad (1)$$

$$\boldsymbol{\eta}\left(O^i, O^j\right) = \begin{cases} 0 & \text{if } i = 0; \\ 1 - \frac{\min(A,B)}{\max(A,B)+\epsilon} & \text{otherwise.} \end{cases} \quad (2)$$

$$\boldsymbol{\varphi}\left(O^i\right) = \max\left(0, \frac{\left\|l(O^i) - l(R^i)\right\|_2}{\max(\left\|l(O^i)\right\|_2, \left\|l(R^i)\right\|_2)} - \theta_t\right), \quad (3)$$

where $A = \left\|l(O^i) - l(O^j)\right\|_2$, $B = \left\|l(R^i) - l(R^j)\right\|_2$, $i \neq j$, $\|\cdot\|_2$ is the L_2 norm, O and R denote the occurrences of the target and related object classes respectively. The function $c(\cdot)$ represents a correlation measurement for feature vector $v(\cdot)$, and $l(\cdot)$ denotes the location of an object. The constant ϵ is for avoiding divisions by zero and $\theta_t \in [0, 1)$ is the distance constraint between the target and related objects. Finally, the problem of finding the best path for the "real" object O^* can be formalized into an optimization problem by including the three constrains as

$$\begin{aligned} &\text{Minimize } \Omega(O^1, O^2, \ldots, O^T) \\ &= \sum_{i=1}^{T} \left\{ \begin{array}{l} \gamma_1 \boldsymbol{\rho}(O^{i-1}, O^i) \\ \quad + \ \gamma_2[1 - P(O^i|C)] \\ \quad\quad + \ \gamma_3[1 - \boldsymbol{\eta}(O^{i-1}, O^i)] \\ \quad\quad\quad + \ (1 - \gamma_1 - \gamma_2 - \gamma_3)\boldsymbol{\varphi}(O^i) \end{array} \right\} \end{aligned} \quad (4)$$

where $\gamma_1, \gamma_2, \gamma_3$ are weighting factors such that $\gamma_1 + \gamma_2 + \gamma_3 = 1$, and T is the total number of occurrences of target object class, and $P(O^i|C)$ is the probability of object occurrence O^i being in the target class C. The optimal solution of this optimization problem can be solved via a dynamic programming procedure assuming the selection of current target object is independent from the previously selected objects.

2) Pre-processing: As aforementioned, with unconstrained lighting conditions and video recording environment, even the same object in different videos may appear in a variety of poses, colors, occluding situations and so on. Besides, the video quality would be another concern for effective object retrieval. Therefore a necessary preprecessing procedure is required for the bounding box image containing the detected object. The pre-processing includes two steps, where the first step is to perform histogram equalization and the second step is to carry out image fusion. (1) **Equalization:** The purpose of equalization is to adjust the global contrast of an image for enhancing the bone structure in the image and reveal more details. Since we target at color images, the operation is applied to the luminance channel in the HSV color space. (2) **Image Fusion:** One disadvantage of the equalization operation is that it will also enhance the contrast of background content hence introduce unnecessary noise, so the fusion step is to balance between the image quality and global contrast level, where the original bounding box image and the equalized image are taken as the two input sources for image fusion. The fusion strategy is the pixel-wise weighted averaging.

3) Object extraction via GrabCut: The object extraction from bounding box image is based on the popular GrabCut algorithm [4]. Different from the traditional GrabCut approach which requires human interaction to provide initial

bounding box for interested object and refine segmentation results, we automate the object extraction procedure without user intervention by taking advantage of the object detection results in the following ways: (1) feeding as input the pre-processed bounding box image with user interested object; and (2) initializing the segmentation process by assigning boundary pixels to background.

B. Object-level Feature Extraction and Similarity Fusion

Traditional color histograms are built on the statistical distribution of image pixels without considering any spatial information, which would fail to distinguish two images with the same color distribution but totally different semantics. To tackle this problem, the auto color correlogram (ACC) algorithm [14] is proposed, which takes into consideration both spatial and statistical information, being able to describe embedded object-level concept in a better way.

CEDD is a popular compact composite descriptor which combines both color and texture features in one histogram [15]. The CEDD histogram is composed of 6x24=144 regions, where the 6 regions are determined by the texture unit and the 24 regions are originated from the color unit. After normalization and quantization, the size of CEDD histogram is limited to 54 bytes per image, making this descriptor suitable for use in large image databases.

To effectively utilize the object segmentation results, we propose to perform object-level feature extraction for both ACC and CEDD features. Specifically, we apply an importance weight to each of the foreground (w_{F}) and background (w_{B}) pixels and obtain the final fused feature vector, where $w_{\mathrm{F}} + w_{\mathrm{B}} = 1$, $w_{\mathrm{F}} \in (0, 1]$, $w_{\mathrm{B}} \in [0, 1)$. It is worth mentioning that the determination of w_{F} and w_{B} are application dependent. For example, under the unconstrained video condition, w_{B} should be minimized to diminish the effect of noisy background; however if the interested object (e.g., an horse) is highly related with background (e.g., grass), hence w_{B} should be increased.

The ACC feature similarity is calculated based on normalized Manhattan distance and the CEDD feature similarity is measured by Tanimoto coefficient. Finally, the similarity score is determined by

$$Sim_F = \lambda_1 \cdot Sim_{ACC} + \lambda_2 \cdot Sim_{CEDD}, \tag{5}$$

where $\lambda_1, \lambda_2 \in [0, 1]$ are the corresponding weights for each type of feature with $\lambda_1 + \lambda_2 = 1$.

Given the fused similarity scores of the query example with each of the items in the database, the retrieval is simply by ranking all the items according to the similarity scores.

IV. Experimental Results

To evaluate the effectiveness of the proposed framework, two set of experiments are conducted. First, we evaluate the performance of our proposed object-level spatial color and texture information integration scheme using benchmark data set and provide the comparison with other state-of-the-art algorithms. Second, we evaluate the effectiveness of the whole video object retrieval framework over real-world data set. The evaluation criteria used in this paper is Mean Average Precision (MAP). The experimental results demonstrate the efficacy of our proposed framework.

A. Evaluation on Multi-layer Information Integration Scheme

The WANG database is a subset of 1000 carefully selected images from the Corel stock photo database. It includes ten classes with 100 images per each category. We first evaluate the performance of individual low-level features such as color and texture as baseline. Shape features are not included because they highly rely on object segmentation results whose performance is not guaranteed in most real-world scenario. In addition, we do not conduct experiment on BOW-based features; however the comparison with those methods are given. The results are illustrated in Figure 1. There are some observations from the figure: (1) color-based features (e.g., AutoColorCorrelogram (ACC), ColorHistogram, JointHistogram, ColorLayout, DominantColor, ScalableColor) outperform texture-based features (e.g., Haralick, Tamura, Gabor); (2) compact composite features (e.g., JCD, CEDD, FCTH) outperform single-channel features; (3) ACC and CEDD features perform the best among all features, which inspired us to explore the integration of these two features.

Based on the experimental observations and analysis, we further conduct experiments to validate the proposed multi-layer object-level spatial color and texture information fusion strategy. Specifically, the foreground and background pixels are assigned equal weights (i.e., $w_{\mathrm{F}} = w_{\mathrm{B}} = 0.5$) since they are considered equally important for natural static images. At the same time, we tune the weight of ACC feature (λ_1) from 1 to 0 with step 0.1 and CEDD feature weight (λ_2) changing accordingly, and observe the respective performance. The results indicate that the fused features outperform the original features with an average of 5% to 10% gain on the MAP values, where the best performance is achieved when $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$.

Finally, we compare the performance of our proposed multi-layer fusion algorithm with the other state-of-the-art algorithms and the results are given in Table I, where the basic features used in each compared algorithm is also listed. The experimental results demonstrate the advantage of our approach over the other existing methods with a 5% to 16% increase on the MAP@100 value.

B. Evaluation on Video Object Retrieval Framework

To demonstrate the effectiveness of the proposed object retrieval framework, a real-world data set is composed (In this experiment, "bag" is taken as an object example due to its popularity. Generally the proposed framework

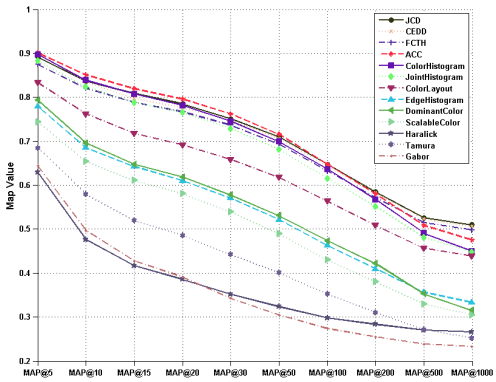


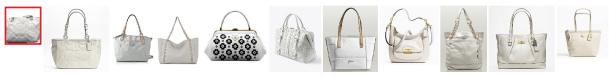
Figure 1: MAP values for low-level individual features on WANG database.

Table I: Feature composition of the state-of-the-art algorithms.

Methods	Features						MAP
	Color	Texture	Shape	SIFT	HOG	LBP	
Hiremath [16]	✓	✓	✓	-	-	-	54.9%
Wang [17]	✓	✓	✓	-	-	-	59.2%
Jurie [18]	-	-	-	✓	-	-	61.7%
Yang [19]	-	-	-	✓	-	-	64.1%
Yu [20]	-	-	-	✓	✓	✓	65.7%
Proposed	✓	✓	-	-	-	-	**70.6%**

applies to an arbitrary object). The data set contains a real-time recorded video and a set of manually-collected images with 371 bags. The experiment targets at retrieving the most similar bags to the ones appeared in the video. The video first goes through the automatic object detection and extraction module, obtaining the detected bags with bounding boxes. Then the bounding box images are applied with the object-level information extraction and integration for final retrieval. Figure 2 (a) displays the retrieval results before applying our proposed information integration strategy, where the leftmost image in red rectangle is the original bounding box image; and (b) shows the results after object segmentation. Apparently the visual results verify the efficacy of our method.

(a) Before object segmentation



(b) After object segmentation

Figure 2: Video object retrieval results.

V. Conclusion

A novel approach for video object retrieval with complex background is proposed in this work. The proposed method is able to automatically extract human interested objects from complex background based on an auto-initiated segmentation algorithm. Spatial color and texture information is then seamlessly integrated and fused together, generating object-level features. Finally, the fused similarity based on different sources of features is obtained for efficient and effective visual retrieval. It is worth mentioning that the proposed multi-layer object-level information integration strategy is applicable to both tasks of image retrieval and image classification. However the efficiency and complexity should be further studied on larger data set in future.

References

[1] Y. Yang, H.-Y. Ha, F. C. Fleites, and S.-C. Chen, "A multimedia semantic retrieval mobile system based on hidden coherent feature groups," *IEEE Multimedia*, p. 1, 2013.

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

[3] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3241–3248.

[4] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3. ACM, 2004, pp. 309–314.

[5] S. Bagon, O. Boiman, and M. Irani, "What is a good image segment? a unified approach to segment extraction," in *Computer Vision–ECCV 2008*. Springer, 2008, pp. 30–44.

[6] Z. Ji, J. Wang, Y. Su, Z. Song, and S. Xing, "Balance between object and background: object enhanced features for scene image classification," *Neurocomputing*, 2013.

[7] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2006, pp. 2169–2178.

[8] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3360–3367.

[9] C. Fredembach, M. Schroder, and S. Susstrunk, "Eigenregions for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 12, pp. 1645–1649, 2004.

[10] H. Cheng and R. Wang, "Semantic modeling of natural scenes based on contextual bayesian networks," *Pattern Recognition*, vol. 43, no. 12, pp. 4042–4054, 2010.

[11] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[12] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *International Conference on Computer Vision (ICCV)*. IEEE, 2009, pp. 606–613.

[13] F. C. Fleites and H. Wang, "Object detection in unconstrained video sequences using multimodal cues," *TCL Research America Technical Report*, 2013.

[14] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Image indexing using color correlograms," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 1997, pp. 762–768.

[15] S. Chatzichristofis and Y. Boutalis, "Cedd color and edge directivity descriptor a compact descriptor for image indexing and retrieval," *Computer Vision Systems*, pp. 312–322, 2008.

[16] P. Hiremath and J. Pujari, "Content based image retrieval using color, texture and shape features," in *International Conference on Advanced Computing and Communications (ADCOM)*. IEEE, 2007, pp. 780–784.

[17] X.-Y. Wang, Y.-J. Yu, and H.-Y. Yang, "An effective image retrieval scheme using color, texture and shape features," *Computer Standards & Interfaces*, vol. 33, no. 1, pp. 59–68, 2011.

[18] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *International Conference on Computer Vision (ICCV)*, vol. 1. IEEE, 2005, pp. 604–610.

[19] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 1794–1801.

[20] J. Yu, Z. Qin, T. Wan, and X. Zhang, "Feature integration analysis of bag-of-features model for image retrieval," *Neurocomputing*, 2013.