Received December 15, 2017, accepted January 18, 2018, date of publication January 23, 2018, date of current version February 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2796638

MindCamera: Interactive Sketch-Based Image Retrieval and Synthesis

JINGYU WANG, YU ZHAO, QI QI, QIMING HUO, JIAN ZOU, CE GE, AND JIANXIN LIAO
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Yu Zhao (zhaoyu@ebupt.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61471063, Grant 61671079, Grant 61771068, Grant 61372120, and Grant 61421061, and in part by the Beijing Municipal Natural Science Foundation under Grant 4152039.

ABSTRACT Composing a realistic picture according to the mind is tough work for most people. It is not only a complex operation but also a creation process from nonexistence to existence. Therefore, the core of this problem is to provide rich existing materials for stitching. We present an interactive sketch-based image retrieval and synthesis system, MindCamera. Compared with existing methods, it can use images of daily scenes as the dataset and proposes a sketch-based image of a scene retrieval model. Furthermore, MindCamera can blend the target object in the gradient domain to avoid the visible seam, and it introduces alpha matting to realize real-time foreground object extraction and composition. Experiments verify that our retrieval model has higher precision and provides more reasonable and richer materials for users. The practical usage demonstrates that MindCamera allows the interactive creation of complex images, and its final compositing results are natural and realistic.

INDEX TERMS Image retrieval, image segmentation, image fusion.

I. INTRODUCTION

With the explosive growth of social media images, the demand of people to edit pictures is increasing. Some easy-to-use image processing applications have become very hot. However, there is no application that provides convenient natural image synthesis according to the mind of users so far. To realize a practical image synthesis system, two main problems should be resolved. One is how to access image materials in our mind quickly. The other one is how to segment the object from the image and blend it seamlessly into a specific background.

Accessing image materials in our mind is mainly related to the field of image retrieval, including text-based image retrieval (TBIR), content-based image retrieval (CBIR) and sketch-based image retrieval (SBIR). CBIR is unusable without sample pictures in this application, and it is difficult to access images that contain the object of a specific shape only based on TBIR. Moreover, the current SBIR systems always process icon images or scenes, without natural image retrieval based on a sketch. When using complex daily scenes as the dataset, contour extraction is an important step due to the inherent domain gap between sketches and photos. The current contour extraction algorithms cannot filter out the background and texture effectively, which limits the development of SBIR. On the other hand, image synthesis refers to the fields of image segmentation and image fusion. Extracting the foreground object from the retrieved image and composing it into the background seamlessly is also a key point for image synthesis by the mind.

To realize image synthesis according to the mind of a user, we develop an interactive sketch-based image retrieval and synthesis system, MindCamera, which uses complex daily scenes as the dataset. The users only need to draw a sketch of an object for retrieval, and our system returns images of scenes that contain the object. Therefore, the users can access other materials quickly, which are likely to appear in the final synthesis. We adopt a novel method of contour extraction to filter out backgrounds and textures, which help form a high-quality line drawings dataset. Furthermore, we use Gradient Field HOG (GF-HOG) [1] to add spatial information to Bag of Visual Word (BoVW) as a descriptor of the sketch. Finally, a feedback of the sorted result to incorporate semantics is provided to improve precision. We also provide text-based Internet image searches to help users find more images.

For image synthesis, two options to extract and compose the foreground object are proposed. The first option is that Grabcut [2] makes a hard segment between the foreground object and the background. Then, Poisson image editing [3] blends target objects in a gradient domain, which makes the synthesis more natural. The other option is that

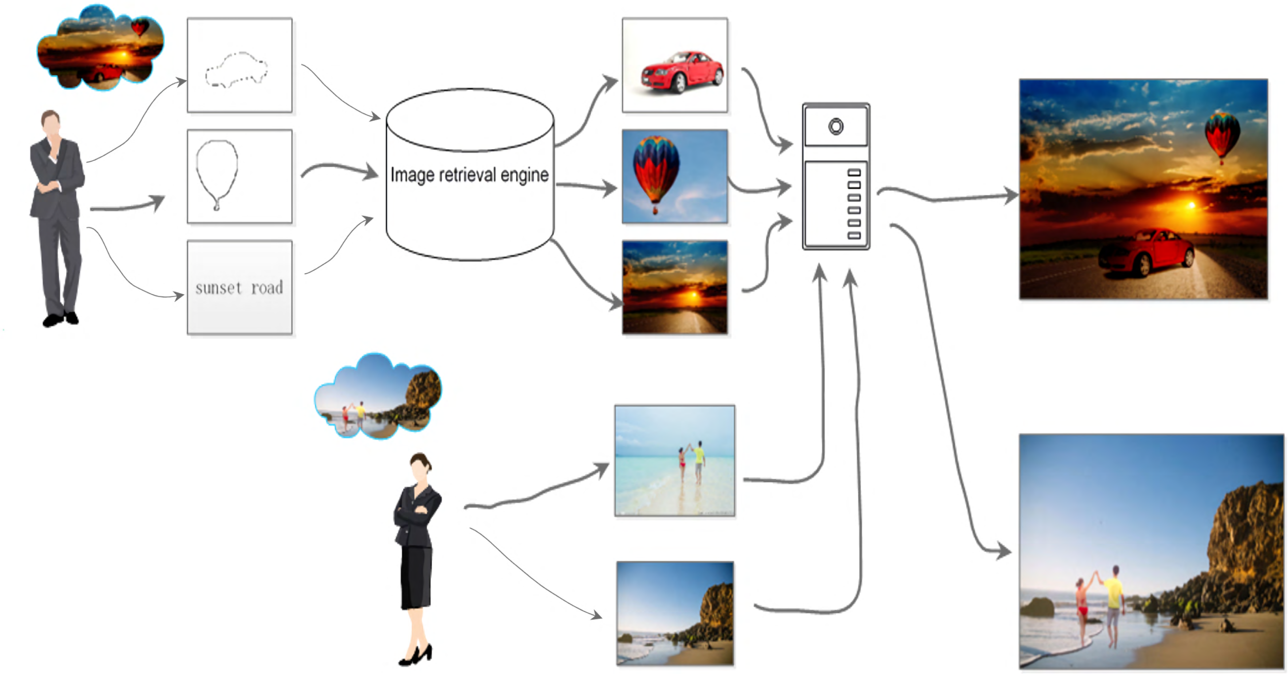


FIGURE 1. **MindCamera: interactive image retrieval and synthesis in large image collection.**

shared matting [4] makes a real-time extraction and composition of foreground objects, which generates a seamless synthesis. Fig. 1 shows several application scenarios of our MindCamera.

The main contributions of this paper are three-fold as follows.

- A whole interactive system that synthesizes images according to the mind of a user is proposed. We combine many optional methods to obtain rich materials, such as handcraft SBIR, text semantics description and upload images, and provide natural image synthesis.
- Natural images rather than icon images are first used as the dataset for SBIR in MindCamera. We design a retrieval model, which makes it possible to retrieve complex scenes with a sketch.
- The proposed contour extraction algorithm can be extended to other SBIR applications to form a high-quality line drawings dataset.

The remaining parts of this paper are organized as follows. Section II briefly summaries the related work. The details of SBIR are given in Section III. Section IV focuses on the implementation of image synthesis. In Section V, a series of experiments are presented to prove the effectiveness of the proposed methods. More discussions and conclusions about the system are presented in Section VI.

II. RELATED WORK

A. SKETCH-BASED IMAGE RETRIEVAL

Recent research on SBIR has been always focused on how to represent sketch features. The dimension of the sparse sketch can be reduced through the feature representation, and the sketch can be represented by its features. We summarize the method of feature representation in the following three ways.

Stroke description-based sketch feature representation [5] processes the stroke as the basic unit and represents the sketch by extracting the feature of the stroke. In 2009, Microsoft developed an early large-scale sketch-based image retrieval system, MindFinder [5], which combines the coordinates and orientations of the edge pixels as the feature and forms a dictionary. In this way, it does not only simplify the shape feature description but also saves the spatial information of the sketch. Finally, it achieves an accurate and rapid retrieval effect through the inverted index. However, this method has strict restrictions on the scale and position, which limits its application.

Combinatorial primitives-based sketch feature representation [6], [7] can represent a complex sketch, since it suggests that the pattern the user wants to retrieve can be made up of one or more primitives. In 2014, Yang *et al.* [6] proposed a method that matches contours by their segments. The general process is to divide the contour into segments at the end of its skeleton. Then, a 12-dimensional feature vector is used to represent each segment, which offers the scale and orientation invariance. All of the feature vectors are arranged clockwise to form a vector group. A similarity matrix of two vector groups can be obtained according to the similarity of two vectors. Then, the Hungarian algorithm is used to calculate the maximum matching, and the result is regarded as the similarity of these two contours. Xiao *et al.* [7] divide the contour into shape words, including the straight line and the curve. They extract the shape word of the contour and then use Chamfer matching to calculate the similarity of the shape words. The inverted index is built to speed up the process. It solves the problem of large memory consumption in MindFinder [8], and its precision is greater than other methods.

Shape feature-based sketch feature representation [9], [10], [11] extracts the global or local features to represent the sketch. Boubekeur and Alexa [9] explore standard HOG within a BoVW framework, computing HOG at random points of the sketch. However, it is difficult to obtain a good result due to the lack of spatial information in BoVW. To improve the performance of HOG in SBIR, Hu *et al.* [10] propose the GF-HOG descriptor algorithm. The orientation of non-contour pixels is interpolated from the orientations of nearby contours under the Laplacian smoothness constraint, which adds the spatial information to the descriptor. Then, the point of the contour is regarded as the point of interest, and HOG is used in different scales. This method solves the shortcomings of lacking spatial information in BoVW and is not sensitive to position, scale and orientation. However, it is sometimes difficult to distinguish different objects of similar shape. Therefore, the semantics of the sketch are often incorporated to achieve a better result.

Although there are many methods of representing the feature of a sketch, the precision of SBIR is still too low for commercial use. We observe that most studies focus on feature extraction, but few have deeply work on contour extraction. Most existing SBIR systems use the Canny algorithm [16] during contour extraction. However, this algorithm cannot filter out the background and texture details. Therefore, it is difficult to form a high-quality line picture collection, which is a major reason for the low precision of SBIR. In addition, the previous work mainly processes iconic object images, and no research has been done on natural daily images. Based on the above problems, we use a novel method of contour extraction, which can filter out the influence of background and texture details effectively. At the same time, we propose a sketch-based image of a scene retrieval model, which makes it possible to retrieve complex scenes with the sketch.

B. IMAGE SYNTHESIS

Chen *et al.* [12] propose a system for automatic image synthesis with keyword-annotated sketches, Sketch2Photo. In this system, keywords trigger a Google image search, which returns possible images. Then, the returned results are chosen through a set of filtering, which helps achieve a good composite result. Sketch2Photo has two main characteristics. First, the retrieval of Internet pictures relies heavily on textual information. Results from the search engine always have a target object in the salient region, which is obviously partial to the following segmentation algorithm. Second, the number of pictures returned by the search engine is large, and the process of filtering these pictures online is cumbersome. In general, the synthesis of a picture takes approximately 30 min, which is obviously unacceptable in practical application. Johnson *et al.* [13] propose CG2Real, which replaces scenes in computer games with real scenes. It makes the picture more realistic while maintaining the same scene structure. In 2011, Wang *et al.* [14] proposed a cartoon picture synthesis tool, Sketch2Carton, which is based on MindFinder [5], and it achieves real-time cartoon synthesis. Eitz *et al.* [15] present an interactive sketch synthesis system called Photosketcher, which does not depend on text, and its speed is fast with preprocessed image datasets. However, the method of feature representation used in Photosketcher offers limited invariance to changes in position, orientation and scale. It is difficult to achieve satisfactory results by only relying on the sketch and lacking the semantics.

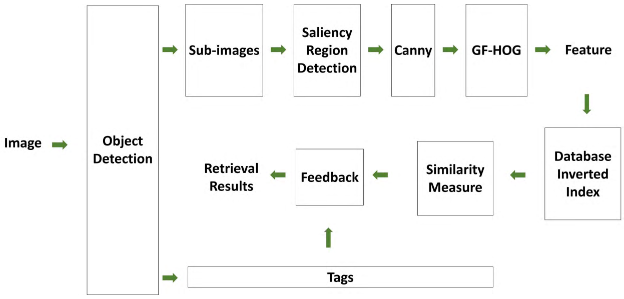


FIGURE 2. **Framework of the proposed SBIR.**

We find that most sketch-based image synthesis systems use iconic object images as the dataset. Researchers have ignored the relationship between different objects. However, related objects are likely to appear together in the suppositional scene of the user. For example, a jumping dog is very likely to appear in the same scene as a frisbee. Accordingly, when the user draws a jumping dog, it would greatly reduce the retrieval time if the frisbee was also in the picture returned by the system. Our sketch-based image of scene retrieval provides the possibility for this case. None of the previously reported image synthesis systems have this functionality.

III. SKETCH-BASED IMAGE OF SCENE RETRIEVAL

To provide users with faster retrieval speed, fewer retrieval times and richer materials, a new sketch-based image of scene retrieval modeling is proposed, depicted in Fig. 2. The proposed SBIR can be used to process daily complex images and return images of scenes containing the target object.

A. OBJECT DETECTION AND CONTOUR EXTRACTION

Contour extraction is a necessary part of SBIR, which breaks up the domain gap between the sketch and the photo. On the one hand, it is difficult to match the right result with the incomplete contour. On the other hand, too many details add interference to the matching because users tend to ignore some details when they draw the sketch. Therefore, a good contour extraction algorithm is vital to SBIR. Classical edge detection algorithms, such as Canny edge detection [16], mostly define the region with large gradient changes as the contour. Canny edge detection is simple and efficient, but its result is not particularly ideal. It cannot distinguish contour from texture. Arbelaez *et al.* [17] propose a contour extraction algorithm, gPb, which takes the color, brightness, and texture

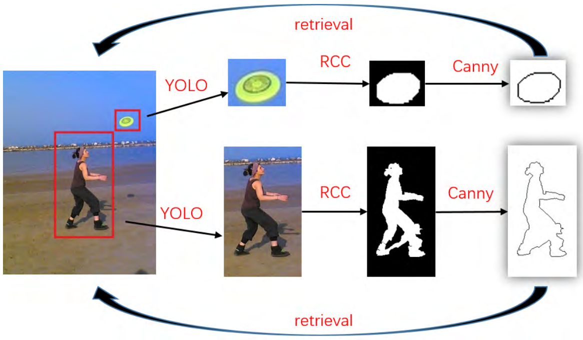


FIGURE 3. Contour extraction workflow.

differences into account. The algorithm can overcome the shortcomings of the Canny algorithm effectively. However, it is quite computationally intensive since it extracts multiscale local gradient features and involves solving a generalized Eigenproblem. Furthermore, there is currently no effective way to remove the background factor during contour extraction.

Our search target is a specific object, which appears in our mind. Accordingly, we use the object detection algorithm to detect objects in the image. The detected object can be cut out through the coordinates of the box around it, and the image is cut into patches containing a single object, i.e., sub-images. At the same time, the mapping relation between sub-images and their original image is recorded. Then, we use the salient region detection algorithm to segment the object and the background. We extract the salient region for each sub-image and generate a binary image with only 0 (background) and 1 (foreground). It is easy for the Canny algorithm to extract the contour from the binary image. The next steps in our model are similar to traditional methods. The reason why we divide the original image into sub-images is based on the following consideration. In our application, users mainly search for a specific object with a specific shape. What they want is not only an object but also to use the returned result to synthesize a new scene. It is better to return images of scenes, which contain the object that the user sketches so the other objects contained in the result may also be the materials the user wants. Therefore, the results that SBIR returns are original images rather than sub-images, as illustrated in Fig. 3.

As for object detection, the ideal method is to use the method of deep learning. The state-of-the-art object detection algorithm YOLO [18] processes images on Titan X at 40-90 frames per second (FPS) and has a mean average precision (MAP) of 78.6% on VOC and MAP of 44.0% on COCO test-dev. It applies a single neural network to a full image. This network divides the image into regions and predicts bounding boxes and probabilities for each region. These bounding boxes are weighted by the predicted probabilities. During preprocessing, a threshold is set to filter out objects of low probability. Excluding these objects is mainly based on the following considerations. In some cases, these objects are not in the saliency region of the picture, which cannot provide a high resolution for our compositing. In other words, they are the background that should be ignored. In addition, some categories are not included during the training of YOLO, so that objects of these categories cannot be recognized by YOLO. Ignoring these objects can improve the accuracy of tags. In this way, the results will be better when semantics are added in the later feedback. Then, the targets are cut from the original image, and the tag of each sub-image is saved. Each sub-image then contains a single object, and complex daily images are transformed into iconic object images.

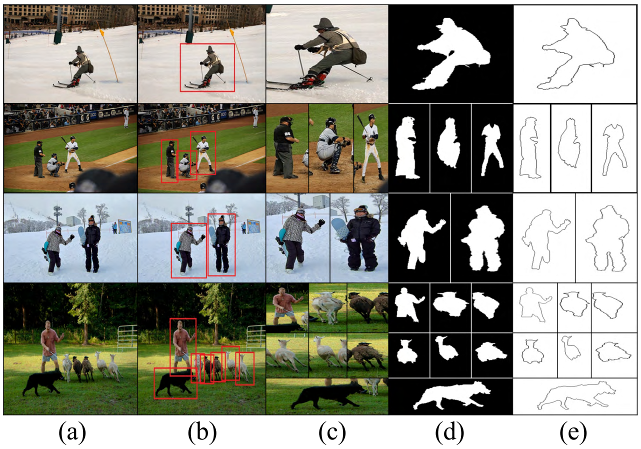


FIGURE 4. Examples of object detection and contour extraction: (a) original image, (b) detected objects marked by a red rectangle, (c) sub-images (d) binary images from SaliencyCut, and (e) contours of sub-images.

The next step is to detect the saliency region of sub-images. In this paper, we introduce SaliencyCut [19] to detect the saliency region of the sub-image. Based on the color contrast histogram, SaliencyCut adds the spatial relation of the pixels and proposes the region contrast histogram. As a result, a saliency map can be obtained and then used to initialize a novel version of Grabcut for a high-quality unsupervised salient object segmentation. A binary image of the sub-image can also be obtained, and its contours can be extracted with the Canny algorithm. Therefore, the background and texture of the sub-image are naturally ignored. It is worth mentioning that SaliencyCut might produce sub-optimal results for images with multiple objects. However, the proposed object detection algorithm overcomes this shortcoming in a way. Fig. 4 shows some examples of the proposed contour extraction.

B. FEATURE REPRESENTATION

BoVW has many advantages, such as simplicity and high efficiency. However, it has an inherent shortcoming in that because the spatial information of feature points is not added in this framework, its accuracy is unsatisfactory. This is also the point that many researchers have focused on. The proposed SBIR adopts BoVW, however, and the spatial information is added during the feature representation.

The GF-HOG descriptor is used here to represent the features of the sketch. This algorithm interpolates the gradient orientation of non-contour points under the condition of Laplacian smoothing constraints. After that, the gradient orientation of these points is strongly affected by the closest points, delivering a form of dynamic neighborhood selection. Therefore, the spatial information is added into BoVW.

First, the binary contour map $M(x, y) = [0, 1]$ is used to calculate the gradient orientation of the contour point. The spare orientation field Ψ is obtained as follows:

$$\theta\left(x, y\right) \to \arctan\left(\frac{\delta M}{\delta y}\Big/\frac{\delta M}{\delta x}\right)\forall_{xy} M\left(x, y\right) = 1 \qquad (1)$$

However, we want to interpolate the gradient orientation of non-contour points and find a dense orientation field Θ_Ω in the entire image coordinates $\Omega \in R^2$. At the same time, Θ_Ω has to keep the gradient orientation of the contour point unchanged: $\Theta(p) = \theta(p)\ \forall_p M(p) = 1$. Therefore, we introduce a Laplace smoothing constraint as follows:

$$\underset{\Theta}{\text{argmin}} \iint_\Omega \|\nabla\Theta - v\|^2 \quad s.t.\Theta|_{\partial\Omega} = \theta|_{\partial\Omega} \qquad (2)$$

On the condition of the Dirichlet boundary [20], the above equation can be solved by the Poisson equation:

$$\Delta\Theta = \text{div } v \text{ over}\Omega \quad s.t.\Theta|_{\partial\Omega} = \theta|_{\partial\Omega} \qquad (3)$$

where $\Delta = \frac{\partial^2}{\partial \text{x}^2} + \frac{\partial^2}{\partial \text{y}^2}$, div is the divergence operation, and $v = \nabla\Psi$, which is the guidance field of the spare orientation field. In the discrete state, this equation can be expressed as the following:

$$\left|N_p\right|\Theta_p - \sum_{q\in N_p\cap q\notin\partial\Omega}\Theta_q = \sum_{q\in N_p\cap\partial\Omega}\theta_q + \sum_{q\in N_p} v_{pq} \qquad (4)$$

where N_p is the four neighbors of the point p, q is one of N_p, and $v_{\mathbf{pq}} = \theta_p - \theta_q$. We can solve this problem with a linear algebraic solution.

Then, multi-scale histograms of the gradient are computed over the gradient field at the point of the contour. Histograms are clustered offline to form a codebook via k-means.

The inverted index makes the linear search of the original complexity O (n), where n is the number of images in the dataset, become the search of the complexity O (k), where k $\ll n$. In this paper, an inverted index is established, which greatly accelerates the retrieval speed of the proposed system and makes it possible for large-scale image retrieval.

C. SIMILARITY MEASURE

The cosine is chosen in this work for the distance measure between two frequency histograms. Combined with the inverse document frequency (IDF), the following equations are given:

$$\text{Sim}\left(\text{Q},\text{D}_i\right) = \frac{1}{M_Q M_{\text{D}_i}} \sum\nolimits_{p\in Q\cap \text{D}_i} f_{Q,p} f_{\text{D}_i,p} IDF_p \qquad (5)$$

$$M_Q = \sqrt{\sum\nolimits_{p\in Q}\left(f_{Q,p}^2\right)} \qquad (6)$$

$$M_{\text{D}_i} = \sqrt{\sum\nolimits_{p\in \text{D}_i}\left(f_{\text{D}_i,p}^2\right)} \qquad (7)$$

$$IDF_p = \ln\frac{N}{f_p} \qquad (8)$$

where Q represents the histogram of the query sketch, D_i represents the images in dataset, N is the number of images in the dataset, f_p is the number of images containing visual word W_p, and $f_{Q,p}$ and $f_{\text{D}_{i,p}}$ are the counts of visual word W_p in the query and image, respectively.

D. INCORPORATING SEMANTICS

Sketch-based queries can specify a target shape, but the semantics of the target are often expressed through the text. Furthermore, the retrieval precision can be improved with the combination of text. Tags of the traditional dataset are accurate, although many human resources are required during data labeling. Different from the previous SBIR systems that incorporate textual information, the tags obtained from YOLO are not accurate. In addition, there are also differences between the description of users and tags of the dataset even for the same object, such as the sea and the ocean. Therefore, it is difficult to search directly through the text in our application.

Considering this, we propose a novel method to incorporate semantics. We find that the proposed SBIR has high precision in the top k when k is small before incorporating semantics, as shown in Fig. 5. According to this, feedback is made to the results by using sub-image tags. Then, the results are partial to objects of the specific category while retaining their shape.

After the sketching, the similarity S_i can be obtained between image i and the sketch, and images are sorted by similarity. The feedback value F_T of the tag T is calculated according to the tags in top k using (9), where C_i is the accuracy of the tag that we get from YOLO, and T_i is the tag of image i. Then, similarities in top n $(n \geq k)$ are recomputed through (10). Finally, images in top n are sorted, and the top k images are returned.

$$F_T = \sum\nolimits_{T_i=T} \frac{S_i}{\sum\limits_{i=1}^{k} S_i} * C_i \qquad (9)$$

$$S_i = S_i^{1-F_T} s.t.T_i = T \qquad (10)$$

In this way, users do not need to enter the text. The system determines the categories that the sketch is most likely to belong to and gives priority to return the most similar objects of these categories. In other words, the more similar the drawing is, the more satisfying the result.

IV. IMAGE SYNTHESIS

In MindCamera, two options for the compositing are provided. One is Grabcut & Possion image editing, which makes the object more natural in some backgrounds. Moreover, in some cases, users do not want to change the whole color of the object, so alpha matting is provided.

A. GRABCUT & POSSION IMAGE EDITING

The classic Grabcut algorithm [2] is used to segment the selected object, which is simple and fast. The user only needs to outline the object, and then the system can segment the

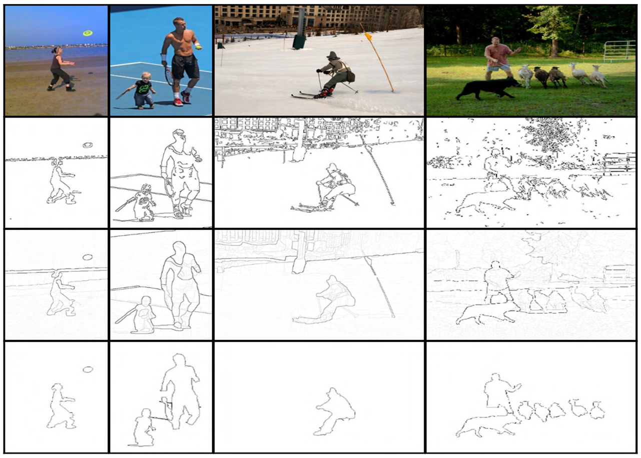


FIGURE 5. Top row: RGB images, Second row: Canny contours, Third row: gPb contours, Bottom row: our contours.

object from the picture. Users are provided with the brush function, which helps edit the object simply and makes the result more satisfying. To make the synthesized image more natural, there should not be obvious seams in the boundary between the source image and the target image. However, when there is a texture difference between the source image and the target image, a direct copy will make an obvious seam. To solve this problem, Possion image editing [3] is used to blend the selected object into the background, which makes the final synthesis natural and realistic. This method considers both the boundary of the source image and the gradient of the target image to find the optimal value of the pixel and, finally, achieves the seamless blending.

B. ALPHA MATTING

To provide more choices, alpha matting is introduced. The matting techniques need to estimate the foreground (F) and background (B) for all pixels of image I, along with opacity (α) values. The relationship between these values can be expressed by the following:

$$C_p = \alpha_p F_p + (1 - \alpha_p) B_p \tag{11}$$

where the observed color $C_p$of pixel p is expressed as a linear combination of F_p and B_p, with parameter α_p.

The Shared Matting [4] algorithm can realize real-time alpha matting. Users only need to draw a boundary curve around the object. The proposed system gets a trimap according to the curve and calculates the alpha matte of the object. Finally, it is fast to copy the object to the background directly. With the alpha matte, it does not only remove visible seam but also keeps the color of the object.

V. EXPERIMENTS AND RESULTS

We compare the proposed method of contour extraction with the traditional algorithms of Canny and gPb. We first extract the sub-image contours and then put the contours together to form a final contour. The result shows that the proposed method filters out the background and textures as well as maintains the basic shape of objects in the picture. This provides high-quality line drawings from the dataset, as shown in Fig. 5.

To evaluate the proposed SBIR, we experiment with the Flickr 160 [10]. The method is compared with the system

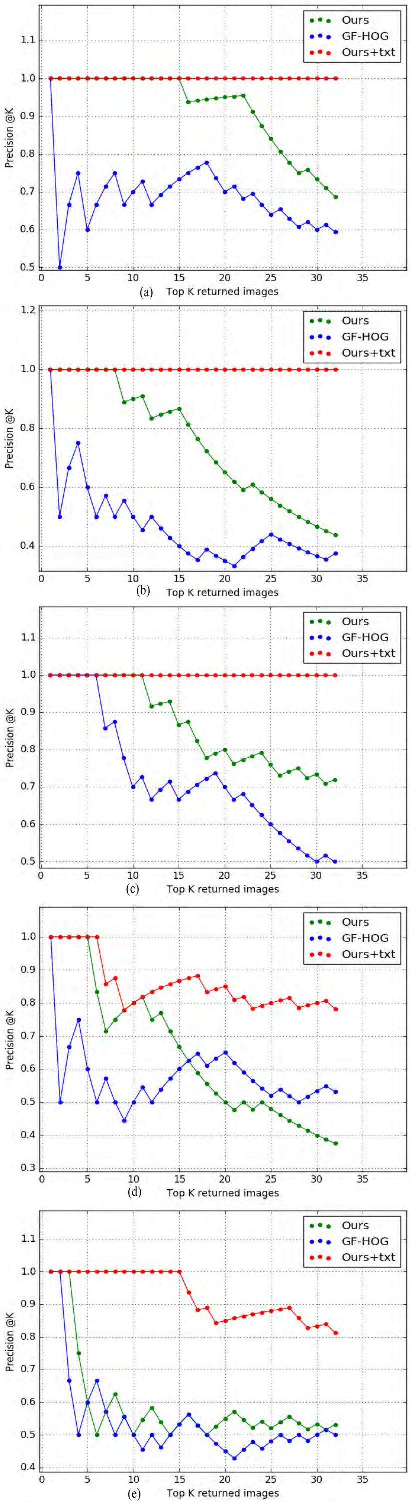


FIGURE 6. Precision@K curves of five examples in Flickr 160 using different methods.

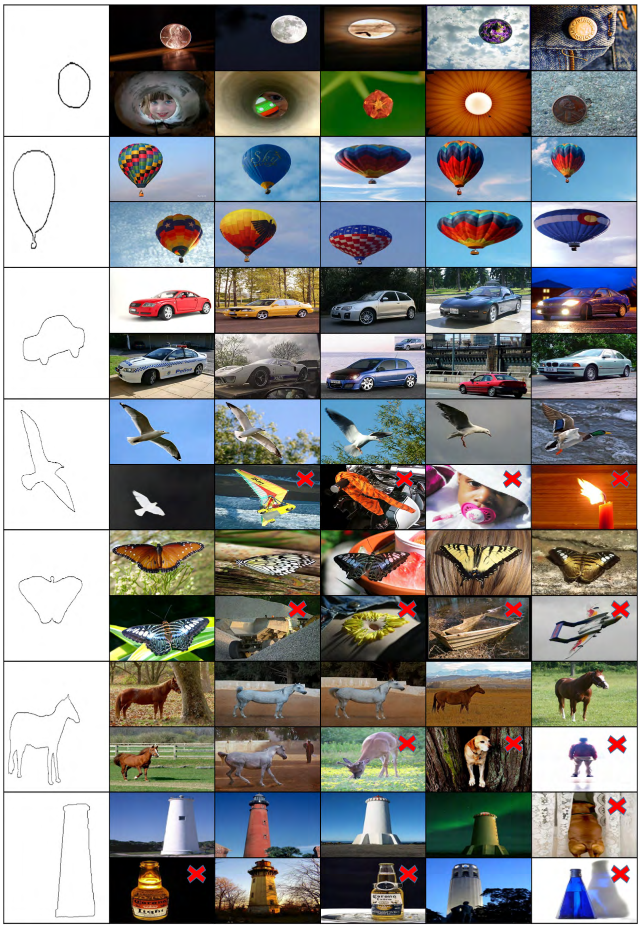


FIGURE 7. Example queries and corresponding Top 10 results for a sketch.

proposed by Hu *et al.* [10], which extracts the contours using the Canny algorithm. The precision of the two methods is compared in the top k. The results are shown in Fig. 6.

It is clear that the average precision of our method is higher than the precision of Rui even without the text feedback. In addition, the proposed SBIR has high precision in the top k, when the k value is small. Giving priority to returning images that meet the need of users is the primary characteristic of our system; this characteristic is used when incorporating semantics. Of course, in some cases, the precision of this method decreases significantly when k is large, even presenting lower precision than the method of Rui. This is mainly because sometimes it is difficult to obtain the complete contour of an object when objects occlude each other. However, this issue is solved when semantics are incorporated. In addition, it can be seen from the figure that the proposed method with the text feedback improves the precision.

This model is tested on the Microsoft COCO validation dataset [21] based on following considerations. First, this dataset is different from previous datasets such as ImageNet, whose images are all iconic object images or iconic scene images. Images in COCO are complex daily scenes containing common objects in their natural context, which is exactly what is needed. When users retrieve an object, SBIR does not only return the object but also other associated objects. Therefore, it provides users with more materials and speeds up the retrieval process. To a certain degree, the contour extraction work is actually the work of manual processing while building the COCO dataset. The COCO verification dataset contains 40,000 daily unclassified images, and many images contain multiple objects. During object detection, these images are cut into 80,000 sub-images that only contain a single object. Fig. 7 shows the sub-images the SBIR returns. Fig. 8 gives some samples after incorporating semantics.

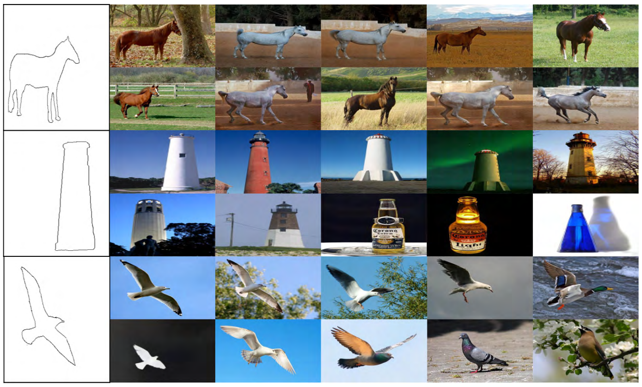


FIGURE 8. Example queries and corresponding Top 10 results for Tag + Sketch compared with Fig. 7.

TABLE 1. Average time at all stages.

Stage	Sketch	Search	G&P	Alpha matting
Time	1–2 min	1–2s	0.5–1 min	0.2–0.5 min

Fig. 9 shows some examples of MindCamera. It can be seen that the results are satisfying. In the final row, a picture in which baseball players are playing baseball on the baseball field is synthesized. We search an image of the baseball field through TBIR as the background. Then, we sketch a batter holding a baseball bat and choose an image that contains many players from the returned images. Other players are needed to appear in the final picture, so we find a pitcher through the same method. Finally, the images are segmented from the original image and blend the segments into the background with appropriate location and size. This SBIR reduces the retrieval time and provides rich materials for the final composition. Finally, the proposed compositing algorithm provides a natural and realistic result.

For a person with simple practice, we give out the average time at all stages when he/she synthesizes a picture, as shown in Table 1. It takes less than 3 min to obtain the result, which is much faster than Sketch2Photo. Moreover, the main time-consuming stage is sketching, where the more similar the drawing is, the more satisfying the result.

VI. DISCUSSION AND CONCLUSION

A. CONCLUSION

We propose an interactive sketch-based image retrieval and synthesis system, MindCamera. This system improves the

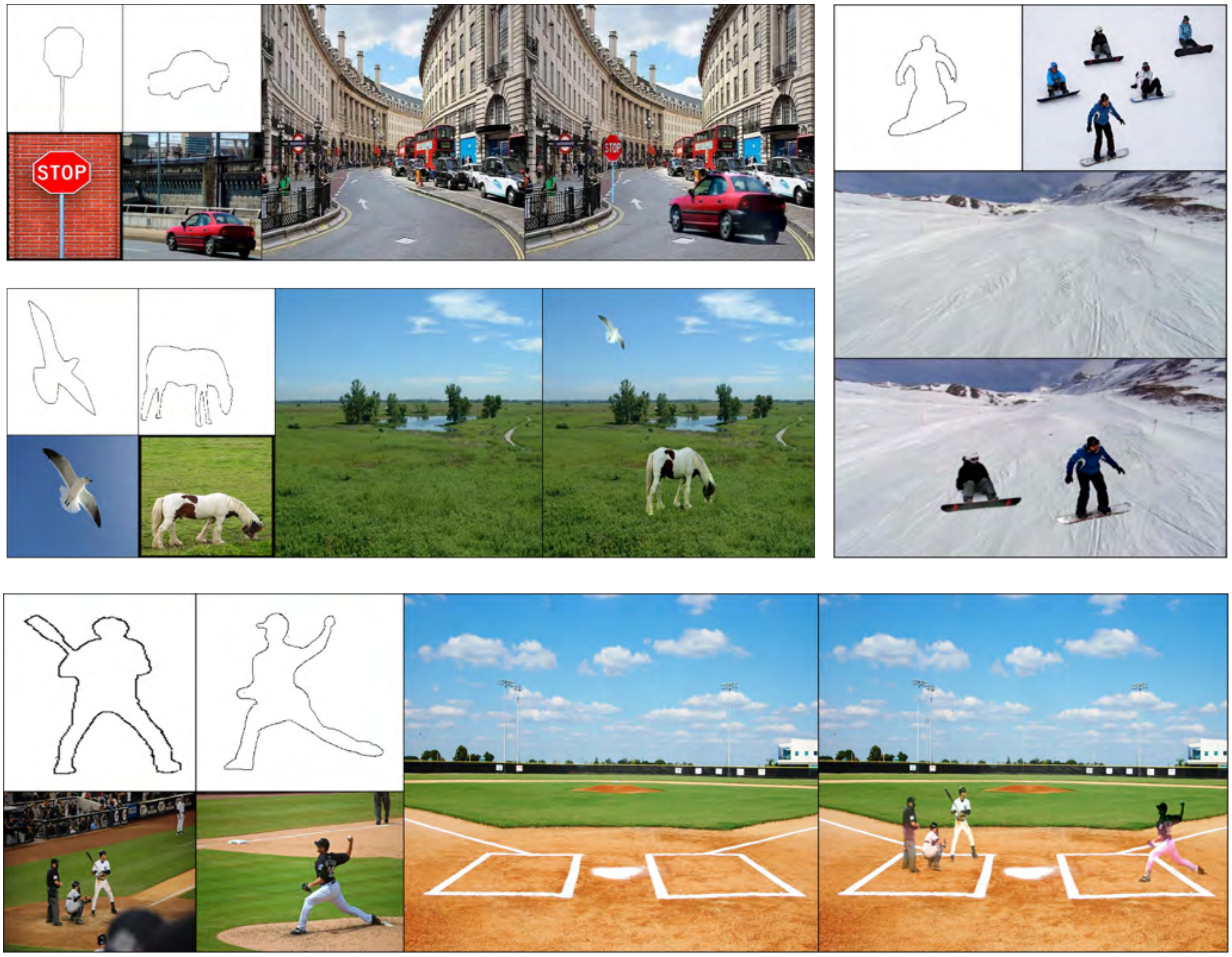


FIGURE 9. Final compositing results created by MindCamera.

usage effect and experience of sketch-based image retrieval and composition based on key contributions. To our knowledge, MindCamera is the first application that processes complex images of daily scenes, and the sketch-based image of scene retrieval provides richer materials for users. This contour extraction algorithm can filter out backgrounds and textures effectively, which helps generate high-quality line drawings. Although the image tags cannot ensure accuracy, the feedback mechanisms are used to incorporate semantics into the image retrieval process. This makes it so the proposed system not only search images but also ''recognizes the sketch.'' Moreover, the composition methods also provide more choices for the user, which contribute to making the final result natural and realistic. The final experiments verify that the proposed method has a good availability and efficiency.

B. FUTURE WORK

In this paper, YOLO is used to transform complex daily images into iconic images, but sometimes segmentation with the bounding box is difficult. Our future work will focus on semantic pixel-level segmentation. In addition, our feedback is affected by the accuracy of tags, so some feature selection theories, such as data representation [22] and the matrix factorization technology [23], [24], will be used to improve the accuracy of tags in the future. Finally, although our contour extraction algorithm filters out the background and texture, it lacks the internal details of an object. When different objects have similar outer contours, it is difficult for our retrieval system to distinguish between these irrelevant objects. Therefore, it is important to add the internal details into the contour, a method that is also worth studying.

REFERENCES

[1] T. Bui and J. Collomosse, ''Scalable sketch-based image retrieval using color gradient features,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV) Workshop*, Dec. 2015, pp. 1012–1019.

[2] C. Rother, V. Kolmogorov, and A. Blake, '''GrabCut': Interactive foreground extraction using iterated graph cuts,'' in *Proc. ACM SIGGRAPH*, 2004, pp. 309–314.

[3] P. Pérez, M. Gangnet, and A. Blake, ''Poisson image editing,'' in *Proc. ACM SIGGRAPH*, Jul. 2003, pp. 313–318.

[4] E. S. L. Gastal and M. M. Oliveira, "Shared sampling for real-time alpha matting," *Comput. Graph. Forum*, vol. 29, no. 2, pp. 575–584, May 2010.

[5] Y. Cao, C. Wang, L. Zhang, and L. Zhang, "Edgel index for large-scale sketch-based image search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 761–768.

[6] C. Yang, O. Tiebe, P. Pietsch, C. Feinen, U. Kelter, and M. Grzegorzek, "Shape-based object retrieval by contour segment matching," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 2202–2206.

[7] C. Xiao, C. Wang, L. Zhang, and L. Zhang, "Sketch-based image retrieval via shape words," in *Proc. ACM Int. Conf. Multimedia Retr. (ICMR)*, 2015, pp. 571–574.

[8] X. Sun, C. Wang, C. Xu, and L. Zhang, "Indexing billions of images for sketch-based retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 233–242.

[9] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 11, pp. 1624–1636, Nov. 2011.

[10] R. Hu, M. Barnard, and J. Collomosse, "Gradient field descriptor for sketch based retrieval and localization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2010, pp. 1025–1028.

[11] P. Xu *et al.*, "Cross-modal subspace learning for fine-grained sketch-based image retrieval," *Neurocomputing*, vol. 278, pp. 75–86, Feb. 2018, doi: 10.1016/j.neucom.2017.05.099.

[12] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2Photo: Internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, Dec. 2009, Art. no. 124.

[13] M. K. Johnson, K. Dale, S. Avidan, H. Pfister, W. T. Freeman, and W. Matusik, "CG2Real: Improving the realism of computer generated images using a large collection of photographs," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 9, pp. 1273–1285, Sep. 2011.

[14] C. Wang, J. Zhang, B. Yang, and L. Zhang, "Sketch2Cartoon: Composing cartoon images by sketching," in *Proc. 19th ACM Int. Conf. Multimedia (MM)*, 2011, pp. 789–790.

[15] M. Eitz, R. Richter, K. Hildebrand, T. Boubekeur, and M. Alexa, "Photosketcher: Interactive sketch-based image synthesis," *IEEE Comput. Graph. Appl.*, vol. 31, no. 6, pp. 56–66, Nov./Dec. 2011.

[16] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[17] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[19] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[20] Z. Ma, J.-H. Xue, A. Leijon, Z.-H. Tan, Z. Yang, and J. Guo, "Decorrelation of neutral vector variables: Theory and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 129–143, Jan. 2018.

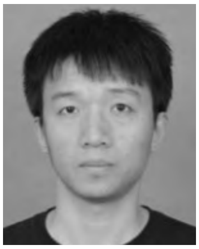
[21] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[22] Z. Ma, Z.-H. Tan, and J. Guo, "Feature selection for neutral vector in EEG signal classification," *Neurocomputing*, vol. 174, pp. 937–945, Jan. 2016.

[23] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 876–889, Apr. 2015.

[24] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2160–2173, Nov. 2011.

JINGYU WANG received the Ph.D. degree from the Beijing University of Posts and Telecommunications in 2008. He is currently an Associate Professor with the Beijing University of Posts and Telecommunications. His research interests include the span broad aspects of SDN, big data processing and transmission technology, overlay networks, multimedia services and communication, and traffic engineering.



YU ZHAO received the B.S. degree from the Beijing University of Posts and Telecommunications in 2016, where he is currently pursuing the M.S. degree. His research interests include image retrieval, machine learning, and image processing.

QI QI received the Ph.D. degree from the Beijing University of Posts and Telecommunications in 2010. She is currently an Associate Professor with the Beijing University of Posts and Telecommunications. Her research interests include cloud computing, consumer electronics, future Internet, ubiquitous services, and reinforcement learning.

QIMING HUO received the B.S. degree from the Beijing University of Posts and Telecommunications in 2016, where he is currently pursuing the M.S. degree. His research interests include artificial intelligence, machine learning, and neural networks.

JIAN ZOU received the B.S. degree from the Guangdong Ocean University. He is currently pursuing the M.S. degree with the Beijing University of Posts and Telecommunications. His research interests include SDN, cloud computing, network traffic analysis, and packet processing. He has participated in several computer contests and received prizes.

CE GE received the B.S. degree from the Beijing University of Posts and Telecommunications in 2016, where he is currently pursuing the Ph.D. degree. His research interests include deep learning, visual recognition, and weakly supervised machine learning.

JIANXIN LIAO received the B.S., M.S., and Ph.D. degrees from the University of Electronics Science and Technology of China in 1985, 1991, and 1996, respectively. He is currently a Professor with the Beijing University of Posts and Telecommunications. He has published over 100 research papers and several books. His main research interests include mobile intelligent networks, service network intelligence, networking architectures, and protocols and multimedia communication.