Visual perception driven collage synthesis

Zuyi Yang¹, Qinghui Dai¹, and Junsong Zhang¹ (✉)

© The Author(s) 2021.

Abstract A collage is a composite artwork made from the spatial layout of multiple pictures on a canvas, collected from the Internet or user photographs. Collages, usually made by skilled artists, involve a complex manual process, especially when searching for component pictures and adjusting their spatial layout to meet artistic requirements. In this paper, we present a visual perception driven method for automatically synthesizing visually pleasing collages. Unlike previous works, we focus on how to design a collage layout which not only provides easy access to the theme of the overall image, but also conforms to human visual perception. To achieve this goal, we formulate the generation of collages as a mapping problem: given a canvas image, first, compute a saliency map for it and a vector field for each sub-region of it. Second, using a divide-and-conquer strategy, generate a series of patch sets from the canvas image, where the salient map and the vector field are used to determine each patch's size and direction respectively. Third, construct a Gestalt-based energy function to choose the most visually pleasing and orderly patch set as the final layout. Finally, using a semantic-color metric, map the picture set to the patch set to generate the final collage. Extensive experimental and user study results show that this method can generate visual pleasing collages.

Keywords collage; gestalt psychology; saliency map; layout optimization; human visual perception

1 Introduction

A collage (see Fig. 1) is a composite artwork made from the spatial layout of multiple pictures on a

canvas, which are collected from the Internet or user photographs. Given a canvas and a set of picture elements, this art form can flexibly assign each element to an appropriate position and size on the canvas, thus creating strong interest and visual appeal. Therefore, it is widely used in design, advertising, rich media creation, and many other decorative illustrations.

However, the manual creation of collages usually requires strong design experience (see Fig. 1) and involves an extensive, tedious process, especially when searching for appropriate pictures and creating a visual appealing layout. Automatic collage synthesis methods also exist, which can be divided into three categories in our experience: (i) traditional collage methods [1–8], focusing on creating a visual and informative summary of a given image set, (ii) picture mosaic methods [9–11], composing multiple sub-pictures to realize a visual simulation of a source image, and (iii) photo montage methods [12, 13], synthesizing a new image from several photographs by cutting, gluing, rearranging, and overlapping operations. However, all of these

Fig. 1 Collages.

¹ Department of Artificial Intelligence, Xiamen University, Xiamen 361005, China. E-mail: Z. Yang, 31520181154411@stu.xmu.edu.cn; Q. Dai, daiqinghui.xingchenwuliao@gmail.com; J. Zhang, zhangjs@xmu.edu.cn (✉).

Manuscript received: 2021-01-18; accepted: 2021-03-14

works fail to take human visual perception into consideration, which our method does. Specifically, compared to previous works, we consider two visual perception mechanisms, the Gestalt principle and visual attention, to ensure that the synthesized collage layout sufficiently conforms to human visual perception. Although some works [1–3, 8] use a visual attention mechanism to extract salient regions of the sub-images to preserve richer information in the collages, they do not use it to ensure thematic relevance between the collage and the canvas image as we do.

Gestalt psychology principles reflect the strategies used by the human visual system to group objects into forms and create internal representations for them. Whenever groups of visual elements have one or several characteristics in common, they are combined to form a new larger visual object [14]. In computer vision and graphics, Gestalt principles have been applied in various ways, e.g., to image or scene abstraction [15, 16], and line or pattern grouping [17–19]. In this paper, we design a Gestalt-based energy function to guide the layout optimization process, which can ensure that the generated patch set is organized in a more reasonable and visual pleasing way. To the best of our knowledge, this method is the first attempt to apply Gestalt principles to the creation of collages.

Our method proceeds as follows: given a canvas image and a set of pictures, we first compute a saliency map for the canvas and a vector field for each of its sub-regions. Secondly, for each region, using the streamlines of the vector field, we adopt a divide-and-conquer strategy to generate several patch sets and use the salience of the location of each patch to control its size. Finally, we use our Gestalt-based energy function to achieve the most visually pleasing and orderly layout, with a novel semantic-color similarity metric to automatically map pictures into patches to generate the final collage.

We have conducted a variety of comparative experiments and user studies to evaluate the effectiveness of this method.

The main contributions of this paper are as follows:

1. A novel system for generating collages. To the best of our knowledge, our method is the first attempt to apply both Gestalt principles and image salience to ensure that the generated collages conform to human visual perception.
2. A novel semantic-color similarity metric to search for suitable pictures jointly considering high-level semantics and low-level color features. To compute the salient map of the canvas, we build an eye movement based visual attention model.

2 Related work

In recent years, many researchers have studied the computer aided design of collages in different contexts. Most focus on creating a visual and informative image summary from a given set of pictures [1–4, 8, 10]. Wang et al. [1] first presented a Bayesian framework to automatically create collages, where the Markov chain Monte Carlo method was used to optimize image arrangement. However, their work does not support layout in arbitrary shapes. Following their work, considering both semantic and visual factors, Battiato et al. [2] designed a self-adaptive image cropping algorithm and used a genetic algorithm to generate the image layout. In addition, Yu et al. [3] developed a collage system with a circle packing method. Furthermore, Liu et al. [8] introduced content correlation between pictures to ensure appropriate proximity, and extracted salient regions of pictures to make full use of the canvas space. Sharing similar ideas, we also consider these two factors to guide the element layout. However, unlike them, we compute the semantic relevance between canvas and pictures, not between pictures, and we generate the saliency map from the canvas, not from the picture elements. Furthermore, the above works only adopt simple shapes, such as a rectangle, circle, or rhombus as the layout canvas, and focus on creating a visual and informative summary from the given image set, to provide a quick and efficient way to browse the image set. Instead, we allow an arbitrary image as the canvas and aim to synthesize a collage which not only conveys the semantic and visual features of the canvas, but also conforms to human visual perception.

Photo montage pioneers another style of collage, where the result is synthesized from two or more images by cutting, gluing, rearranging, and overlapping operations [12, 13]. Goferman et al. [12] presented a framework to produce an informative and pretty photo montage by exactly cutting interesting regions of images in a puzzle-like manner. Huang et al. [13] created an Arcimboldo-like collage from cutouts

of multiple Internet images. They first used a mean shift clustering approach to automatically segment the input image into patches, and then selected a cutout for each patch with a component-aware cutout matching method. Finally, they assembled these cutouts with affine transformations. To some extent, the generative part of our work is similar, because we both aim to find a suitable picture for each patch using a similarity metric. However, their work mainly considers color and shape similarities, while our work adopts a semantic-color metric to select suitable pictures automatically.

Besides photo montage, picture mosaicing [9–11] is also another way of synthesising collages. A source image is periodically divided into tiled sections (usually of equal size) and each section is replaced with one matching photograph. In the work of jigsaw image mosaics [9], Kim and Pellacini divided the given canvas image into several arbitrary-shaped tiles as compactly as possible and optionally deformed them slightly to achieve a more visually pleasing effect. Unlike their method, we use the saliency map to search for more variable-sized patches and generate

the final layout in a discrete way. Most importantly, we introduce Gestalt principles to optimize the layout to achieve a result more consistent with human visual perception.

3 Method

3.1 Overview

An overview of our collage generation method is illustrated in Fig. 2. Our system takes a canvas image and a set of pictures as input, and outputs a synthesized collage. First, we compute a saliency map for the canvas and a smooth vector field for each sub-region of the canvas. Then, along the vector field, we partition each region into a patch set with a divide-and-conquer strategy, using the saliency map to control the patch size. Additionally, by adjusting the control parameters, we may obtain a series of patch sets for each region. Next, by minimizing a Gestalt-based energy function, we determine the most visually pleasing and orderly patch set for the layout. Finally, using a semantic-color similarity metric, we map the pictures to the patches to produce the final collage.

Fig. 2 Overview. Given a canvas image (1, above), we first generate a vector field and a saliency map (1, below). Next, we adopt a divide and conquer strategy to obtain a series of patch sets along the vector field; each patch's size is controlled by the salience of its location. Then, based on Gestalt principles (2, above), we use an energy function to choose the best patch set (2, below). Finally, using a semantic-color similarity metric, we map the picture set to the patch set to generate the final collage (3, below).

3.2 Vector fields

The first step of our method is to design a desirable vector field for each sub-region of the canvas, which is used to guide the direction of the pictures. The vector field should preserve both the regional and textural features of the canvas. We calculate the vector field using the following steps: first, we decompose the canvas into several regions based on an edge detection algorithm. Then, we apply Delaunay triangulation to further divide each region into a triangular mesh. Finally, we interactively set heat source points in the mesh and record the heat diffusion direction in each triangle as the direction of the vector field. Then we sample a series of streamlines from the vector field to guide patch directions during layout.

3.3 Saliency map

Before partitioning each region, we need to compute a saliency map for the canvas, which will be used to decide where to place bigger or smaller patches. The saliency map measures the likelihood that each position in the canvas attracts the attention of a human observer.

Inspired by Judd et al. [20], we design an eye movement based visual attention model to obtain the saliency map. Specifically, we adopt the MIT saliency database [21] as training data. The database contains a source image set; each source image has one set of corresponding eye movement data. Since the original eye movement data is represented as a series of discrete eye movement points on the map, we perform Gaussian convolution on them to generate a continuous saliency map and label the obtained maps as the training ground truth. Additionally, we also extract a series of feature maps for each source image, including facial and low-level features. The training steps are as follows: first, select 100 pictures from the database and further divide them into two parts, training set (90%) and test set (10%). Second, obtain samples from the selected images. Specifically, we extract 100 positive samples and 100 negative samples from the top 30% and bottom 50% salient areas of each image respectively, with 20,000 samples in total. Third, we compute the sample feature. In particular, for each sample, we extract the gray pixel values of its location in the feature maps and combine the extracted values into a feature vector. Finally, using the samples, we train a linear

support vector machines (SVM) as a practical model. In the forecasting process, we extract the same feature maps from the given canvas and utilize the trained visual attention model to calculate its saliency map.

3.4 Patch sets

After obtaining the streamlines and the saliency map, we search for a patch set for each region. Region partition is a classical problem, where a set of patches of the same or differing sizes are arranged into the given region. In our case, each patch is treated as a square, whose size is decided by the region area and its location's saliency.

3.4.1 Problem formulation

Given a region $\Omega \in R^2$ and n streamlines $SL = \{sl_i\}_{i=1}^n$, the region partition problem in our case is to compute a best configuration for all the patches $P = \{p_j\}_{j=1}^m$ encompassed in Ω . Furthermore, we must ensure that there is no overlap between patches and the patches are arranged based on the streamlines. In addition, we introduce a coverage rate τ to ensure that all patches approximately cover the whole region. Specifically, region partition is defined by the following optimization problem:

$$\begin{aligned} & \text{Maximize } \sum_{j=1}^m \text{Area}(p_j) \geq \tau \text{Area}(\Omega) \\ & \text{Subject } \begin{cases} p_i \subseteq \Omega, & i \in 1, \dots, m \\ p_i \cap p_j = \emptyset, & i, j \in 1, \dots, m, \quad i \neq j \end{cases} \end{aligned} \quad (1)$$

The arrangement of a patch set $P = \{p_j\}_{j=1}^m$ in R^2 is represented by a configuration $C(L, D, S) = \{(l_1, \dots, l_m), (d_1, \dots, d_m), (s_1, \dots, s_m)\}$, where l_j , d_j , and s_j are the location, direction, and size of the j th patch p_j , respectively. If all patches satisfy the two constraints in Eq. (1), we say that the configuration is valid. Note that each patch is a closed square and thus the patch's area in the first constraint may be calculated based on its size. The inclusion and intersection relations in the second constraint are computed pixelwise.

3.4.2 Finding patches

As mentioned above, we need to arrange the patches along the streamlines, so we traverse the streamlines one by one. In addition, we use a top-down search strategy to maximize the size of patches and minimize the number of patches, which also accelerates the search process. We first consider the patch with the biggest size as cover for the region. If no patch is

found, we decrease the patch size by half and search again. This operation is performed repeatedly until the constraints in Eq. (1) are met.

The above algorithm can obtain a patch set that covers the region with a desired cover rate. However, it may not satisfy the requirement that the patch's size matches its location's saliency: based on the saliency map, a large-sized patch should be placed in a position with higher saliency, and vice versa. Thus, we use a lightness threshold $\xi(s)$ to decide whether a patch with a certain size s can be placed in a position. The lightness threshold $\xi(s)$ is defined as

$$\xi(s) = \text{lightness}_{\max} - \lambda \log_2 \frac{\text{size}_{\max}}{s} \quad (2)$$

where λ is a relaxation factor, which can be adjusted to obtain different lightness thresholds. After finding a patch, we first calculate the average lightness of all pixels in the patch. If the average lightness is greater than the lightness threshold, we think the patch is valid and add it to the patch set P . Otherwise, we abandon the patch and search for a new one along the streamline.

However, a patch obtained by the above strategy may appear to be somewhat visually disordered, because it may intersect visually with an adjacent patch of the same size, coming from other streamlines: see the blue circle in Fig. 3(b). To solve this problem, we selectively traverse the streamlines. The selection

criterion is based on whether more than one patch is found in the current streamline. If so, we divide other streamlines into two sets, namely, the streamlines inside (1) and outside (2) the region surrounded by the extended line of the found patches. Then, we search for patches with half the size along the first set of streamlines, and search for patches with the same size along the second set of streamlines. This divide-and-conquer strategy may be detailed as follows:

1. Sequentially choose a streamline sl_i in the set of streamlines SL .
2. Traverse the current streamline sl_i . If two points sl_i^j and sl_i^k are found meeting $\|sl_i^j, sl_i^k\| = s$ ($1 \leq j < k \leq N$), construct a square patch p and use the line from sl_i^j to sl_i^k as the center line of the patch. Then, go to step 3. Otherwise, go to step 5.
3. If p in Ω and $p \cap P = \emptyset$, calculate the average lightness l of patch p and go to step 4. Otherwise, go to step 2.
4. If $l \geq \xi(s)$, add p to P and update C and go to step 5. Otherwise, go to step 2.
5. If $\sum_{j=1}^m \text{Area}(p_j) \geq \tau \text{Area}(\Omega)$, go to step 8. Otherwise, go to step 6.
6. If streamline sl_i has more than one patch, divide the streamlines into two sets SL_1 and SL_2 . SL_1 is inside the region surrounded by the extended line of the found patches, and SL_2 is outside the region. Go to step 7. Otherwise, go to step 1.
7. Set patch size $s/2$ for SL_1 and s for SL_2 . Go to step 1.
8. Return P and C .

3.5 Modeling Gestalt principles

With the divide-and-conquer algorithm and the saliency map, we obtain a good patch set which approximately covers the whole region. However, due to the complexity of the vector field and the uncertainty of the distribution of different saliency maps, if using a fixed lightness threshold, the visual effect of the patch set may be chaotic. Gestalt psychology principles reflect the strategies of the human visual system used to group objects into forms and create internal representations for them. Whenever groups of visual elements have one or several characteristics in common, they are combined to form a new larger visual object [14]. Thus, we use Gestalt principles to measure the degree to which a patch set meets the expectations of human visual

Fig. 3 Divide-and-conquer strategy. (a) Vector field. Red arrows: direction of the vector field. Blue circles: heat source points. (b) Result without divide-and-conquer strategy. (c) Result with divide-and-conquer strategy.

perception. Specifically, we choose to apply four Gestalt principles to synthesize a collage: proximity, similarity, closure, and regularity. They are defined as follows:

1. *Proximity*: the proximity principle (see Fig. 4(a)) states that when an individual perceives an assortment of objects, those objects close to each other are usually perceived as a group. We measure proximity between two patches p_i and p_j as

$$\text{Proximity}(p_i, p_j) = \|l_i - l_j\| - (s_i + s_j) \quad (3)$$

2. *Similarity*: the similarity principle (see Fig. 4(b)) states that elements within an assortment of objects are usually perceptually grouped together if they are similar to each other in terms of some qualities, such as shape, size, color, orientation, and so on. We measure the similarity of a patch p_i as

$$d_s(p_i, p_j) = \begin{cases} s_i/s_j, & \text{if } s_i \geq s_j \\ s_j/s_i, & \text{otherwise} \end{cases} \quad (4)$$

$$d_d(p_i, p_j) = \frac{d_i \cdot d_j}{|d_i||d_j|} \quad (5)$$

$$\begin{aligned} \text{Similarity}(p_i) = & \omega_{\text{drec}} \sigma \left(\bigcup_{j \in N(p_i)} d_d(p_i, p_j) \right) \\ & + \omega_{\text{size}} \sigma \left(\bigcup_{j \in N(p_i)} d_s(p_i, p_j) \right) \end{aligned} \quad (6)$$

Fig. 4 Gestalt principles. (a) Proximity. (b) Similarity. (c) Closure. (d) Regularity.

3. *Closure*: the closure principle (see Fig. 4(c)) states that when objects are incomplete, an individual usually perceives them as being a whole. We measure the closure of the i th patch p_i as

$$\text{Closure}(p_i) = \frac{d_i \cdot d_j}{|d_i||d_j|} \|l_i - l_j\| \quad (7)$$

4. *Regularity*: the regularity principle (see Fig. 4(d)) states that objects regularly spaced tend to be grouped. We measure the regularity of the i th patch p_i as follows:

$$\text{Regularity}(p_i) = \sigma \left(\bigcup_{j \in N(p_i)} \text{Proximity}(p_i, p_j) \right) \quad (8)$$

Using the above Gestalt principles, we obtain a series of patch sets by adjusting the lightness threshold and choose the most visually pleasing and orderly result by minimizing the following energy:

$$\begin{aligned} E = \min_p \frac{1}{N} \sum_{i=1}^N (\omega_s \text{Similarity}(p_i) \\ + \omega_c \text{Closure}(p_i) + \omega_r \text{Regularity}(p_i)) \end{aligned} \quad (9)$$

where $\omega_s, \omega_c, \omega_r$ are the weights for similarity, closure, and regularity respectively, whose sum is equal to one, and N is the number of patches in set P . To verify the effectiveness of the energy function, we generated a series of visually different patch sets and calculated their energy values, as shown in Fig. 5. It is easy to see that the layouts with smaller energy values are more orderly and visually pleasing.

3.6 Semantic-color mapping

In this step, we need to find a mapping between pictures and patches. Previous works usually choose input pictures to fit the given canvas based on the color of the pictures rather than semantics. We argue that more important pictures should be emphasized by assigning them more space to give a more informative collage. Thus, we apply a novel semantic-color similarity metric to evaluate each picture and choose the most similar picture for each patch. Specifically, a tradeoff exists between the high-level semantics and low-level color features. For high-level semantic features, we would like to choose pictures closely thematically related to the canvas image, whereas, for low-level color features, we prefer to choose pictures with higher color similarity to the corresponding patch. Therefore, our semantic-color similarity metric is defined as follows:

Fig. 5 Different patch sets and their energies.

$$E_{\text{match}}(P_i, I_j) = \omega_{\text{sem}} D_{\text{EJ}}(W(P_i), M(I_j)) + \omega_{\text{col}} D_{\text{INT}}(H(P_i), H(I_j)) \quad (10)$$

where ω_{sem} and ω_{col} are weights. D_{EJ} and D_{INT} measure the semantic and color similarity between the i th patch P_i and the j th picture I_j , respectively, calculated as detailed below.

3.6.1 Semantic similarity

If the picture element's semantics conform to the theme of the canvas, a more informative and visual appealing collage can be synthesized. To achieve this goal, we use the pictures collected by Patterson et al. [22] as our basic picture set; it has 14,340 pictures. In addition, each picture has 102 discriminative attributes, which constitute the probability distribution vector W of the picture. For example, "Trees 1.0", "glass 0.8", and "flowers 0.8" indicate that the picture contains grass, trees, and flowers with 100%, 80%, and 80% probability respectively. Also, we manually assign the canvas image a vector of the same dimension M . Finally, we use Jaccard distance to measure the semantic similarity between the canvas image and the picture element.

$$D_{\text{EJ}}(W, M) = \frac{W \cdot M}{\|W\|^2 + \|M\|^2 - W \cdot M} \quad (11)$$

Note that we directly use the above result to measure the semantic similarity between the patch and the picture, instead of calculating a separate value for each patch.

3.6.2 Color similarity

We measure the color similarity using color statistics from the HSV color histogram. From our experimental experience, we set the hue, saturation and value channels to contain 8, 16, and 4 bins, respectively. Let H_1 and H_2 denote the HSV histograms of a patch and a picture. We compute their histogram intersection distance using:

$$D_{\text{INT}}(H_1, H_2) = \sum_i \min(H_1(i), H_2(i)) \quad (12)$$

where $H(i)$ denotes the proportion of pixels that fall into the i th bin and D_{INT} represents the color similarity.

As mentioned above, we hope for large patches to be assigned a more thematically related picture, as patch size is determined by the saliency of its location. For small-sized patches, we prefer to fit pictures with higher color similarity, which can better preserve the visual features of the canvas. Thus, we use the patch's size to determine the relative weights for semantics and color:

$$\omega_{\text{sem}}(s) = \lambda s / \text{size}_{\text{max}} \quad (13)$$

$$\omega_{\text{col}}(s) = 1 - \omega_{\text{sem}}(s) \quad (14)$$

where s is the current patch size, size_{max} is the default or user-set maximum size, and λ is the tradeoff coefficient.

4 Results and evaluation

4.1 Approach

In this section, we first evaluate the performance of our method according to three criteria:

1. *Image saliency*: whether larger pictures are placed at more salient positions, and vice versa.
2. *Semantic-color similarity metric*: whether picture elements and the canvas are thematically related and whether picture element colors conform to the color of their locations.
3. *Gestalt principles*: whether the use of Gestalt principles contributes to a more visually pleasing and orderly layout.

In addition, we also compare our method with Shape Collage [23] and Arcimboldo-like collage [13], to assess the overall visual effects of our generated collages.

4.2 Criteria evaluation

First, we generated a group of collages using a "dog" image as the canvas to validate whether the saliency

map is useful (see Fig. 6). The salient areas are mainly located on the back, face, and mouth of the dog. Figure 6(d) is a close-up of the region in the red rectangle in Fig. 6(c), containing both salient and non-salient areas. In Fig. 6(d), we can see that we successfully place some big patches in the salient area (blue oval), and some small patches in the non-salient area (red rectangle). Through this strategy, we can preserve the thematic and visual characteristics of the canvas to the maximum extent.

Second, we generated three results with different considerations of semantics and color to validate the semantic-color similarity metric (see Fig. 7). Specifically, we used “flowers” as the semantic attribute to search for suitable pictures from the picture database. Figure 7(a) shows the canvas image, a beautiful flower. Figure 7(b) is the result considering both semantics and color, where $\omega_{\text{sem}} = \omega_{\text{col}} = 0.5$. Figure 7(c) is the result only considering semantics; $\omega_{\text{sem}} = 1$. Figure 7(d) is the result only considering color; $\omega_{\text{col}} = 1$. The constituent elements of Fig. 7(c) are all flower-related pictures. However, some pictures’ color mismatch the color of the canvas. Figure 7(d) has a strong color similarity between the elements and the canvas, but its images are not semantically relevant enough. Figure 7(b) combines the good aspects of both, and shows strong semantic and color similarity, demonstrating the

effectiveness of our semantic-color similarity metric. With this metric, the generated collage achieves a tradeoff between retaining the semantic and visual characteristics of the canvas.

Third, we generated two patch sets with and without Gestalt principles to validate the effectiveness of the Gestalt principles. Figure 8(a) shows the direction of streamlines. Figures 8(b) and 8(c) show results generated with and without Gestalt principles, respectively. Figure 8(c) clearly shows two problems: (i) patch size in some areas varies significantly (red oval), and (ii) the patch layout in some areas is disorganised (blue oval). The Gestalt principle overcomes these problems efficiently. The patch set in Fig. 8(b) looks more orderly and visually pleasing and validates the effectiveness of the Gestalt principles.

Through the above experiments, our system is shown to effectively consider image salience, the semantic-color metric, and Gestalt principles. The saliency map and the semantic-color metric give users a better understanding of the theme and visual characteristics of the canvas image. Meanwhile, the use of Gestalt principles contributes to a more visually pleasing layout and furthermore achieves a collage more consistent with human visual perception.

Fig. 6 Collage with theme “dog”: (a) source image; (b) saliency map; (c) generated collage; (d) detail within red rectangle.

Fig. 7 Semantic-color similarity. (a) Source image. (b) Result generated considering both semantics and color. (c) Result only considering semantics. (d) Result only considering color.

Fig. 8 Use of Gestalt principles. (a) Vector field. Red arrows: direction of the vector field. Blue circles: heat source points. (b) Result generated with Gestalt principles. (c) Result generated without Gestalt principles.

4.3 Comparisons

To verify the overall visual effectiveness of the results, we compared different collages generated by three methods, Shape Collage [23], Arcimboldo-like collage [13], and our method (see Figs. 10 and 11). We also conducted a user study in the form of an online questionnaire to quantitatively evaluate these collages. Because of the differences in these three styles, we only asked users to compare the overall visual effect. We asked twenty participants, eleven female and nine male students with an average age of 22, to score the results of different methods. Each participant graded the pictures on a scale from zero (poor) to seven (excellent), for their opinion on the overall visual effect of the results. Average scores are presented in Fig. 9.

First, we compared our system with Shape Collage [23], popular collage software worldwide. It can create collages with good visual effects in a few seconds and supports arbitrary shapes. More importantly, its results are of a similar kind to ours, as we both arrange the elements in a discrete way.

As shown in Fig. 10, Shape Collage preserves the shape of the canvas well. However, it does not

(a)

(b)

Fig. 9 Average scores of the results of the three methods in the user study.

consider the semantic-color similarity as we do, which causes a disorganized collage layout. On the contrary, our method preserves both the visual and shape features of the canvas well. Besides, it retains more image details than Shape Collage, such as the face of the Mona Lisa, because it can adaptively adjust picture element size based on regions' areas. In the user study, our method obtains generally higher scores than Shape Collage. Also, we applied ANOVA to the collected data and got ($P_{monalisa} = P_{zen} = 0.0$), showing that there are significant differences between the two groups of scores. This comparison shows that our method achieves a better overall visual effect than Shape Collage.

Second, we also compare our method with Arcimboldo-like collage [13]. By filtering large numbers of Internet images, Arcimboldo-like collage combines multiple thematically-related image cutouts to represent the input canvas image. The selected cutouts are purposefully arranged so that the whole assembly represents the input image in both shape and color, while individual cutouts still are recognizable as themselves. Arcimboldo-like collage arranges elements in a continuous way, while our

Fig. 10 Comparison with Shape Collage. Left to right: (a) source image, (b) results of Shape Collage, (c) results of our method.

method adopts a discrete approach, streamline-based arrangement, to lay out the pictures. Therefore, for comparison, we use the original pixels of the canvas to fill the non-laid-out areas, i.e., the gaps between the patches (Fig. 11(c)).

Figure 11 shows a comparison between Arcimboldo-like collage [13] and our method. In Fig. 11(b), multiple meaningful and thematically-related cutouts constitute the visual representation of the source images (Fig. 11(a)), such as a “vegetable” elephant, a “fruit” panda, and a “music” transformer. Our method (see Fig. 11(c)) considers more the semantic and color relevance between the canvas and the elements. Also, we use much smaller elements to visually represent the source images. Furthermore, our method applies two further visual perception factors, a saliency map and Gestalt principles, to guide the layout of the elements. The former is used to retain the thematic features of the canvas, and the latter contributes to a more visually pleasing and orderly layout. From the detailed views of Fig. 11(c), we can see that

some large thematically-related pictures are placed in the salient regions (blue rectangle). Note that though some other regions are also salient, such as the tip of the trunk and edges of the elephant, they are too small to place large-sized pictures without damaging the shape of the canvas. Figure 9(b) gives the scores of these two methods in the user study. Relative scores vary for the three pictures and two methods. The p values obtained from ANOVA are ($P_{\text{elephant}} = 0.00$, $P_{\text{panda}} = 0.86$, $P_{\text{transformer}} = 0.23$). We can thus conclude that these two methods achieve a similar evaluation overall.

4.4 Runtime

Our system is implemented on a PC with a 3.2 GHz CPU and 8 GB memory. The time required for computation depends on vector field generation, layout optimization, and user interaction. If the time consumed in user interaction is ignored, it takes about 30 min to generate a result. More results are shown in Fig. 12.

Fig. 11 Comparison with Arcimboldo-like collage. Left to right: (a) source image, (b) results of Arcimboldo-like collage, (c) results of our system. Detailed views of salient areas of the elephant and panda images are shown on the far right.

Fig. 12 More results.

4.5 Limitations

Our system still has a few limitations. First, it only provides a limited number of types of vector fields. If the canvas image contains many irregular regions, the vector fields generated from these regions may be unexpected, and even against user intent. Second, it is time consuming to generate a collage, particular because of the the process of finding patches and searching for suitable pictures. To address this issue, one potential solution is to use the GPU to speed up the algorithm. Third, our system may produce results with unsatisfactory visual effects, when there are no suitable pictures to satisfy both the semantic and visual features of the canvas. Extending our image database to cover more image themes could help to generate visual pleasing collages in future.

5 Conclusions

In this paper, we have proposed a novel visual perception driven method for creating collages. Given a canvas image and a set of pictures, we first compute a saliency map for the canvas. We then segment the canvas into several regions and calculate a vector field for each. Second, along the vector fields, we search for several patch sets using a divide-and-conquer strategy, using a saliency map to determine patch size. Third, we use a Gestalt-based energy function to achieve the most visually pleasing and orderly layout. Lastly, using a semantic-color metric, we map the pictures to the patches to get the final collage. The collages generated by our method can not only achieve a visual simulation of the canvas, but also enhance the semantic theme of the canvas. More importantly, this method is the first to introduce Gestalt principles into the creation of collages, making the generated results more consistent with human visual perception. We believe that this method is a great demonstration of the combination of cognitive psychology and art computing.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61772440), the Aeronautical Science Foundation of China (No. 20165168007), and Science and Technology of Electro-optic Control Laboratory.

References

- [1] Wang, J. D.; Quan, L.; Sun, J.; Tang, X. O.; Shum, H. Y. Picture collage. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 347–354, 2006.
- [2] Battiato, S.; Ciocca, G.; Gasparini, F.; Puglisi, G.; Schettini, R. Smart photo sticking. In: *Adaptive Multimedia Retrieval: Retrieval, User, and Semantics. Lecture Notes in Computer Science, Vol. 4918*. Boujemaa, N.; Detyniecki, M.; Nürnberger, A. Eds. Springer Berlin Heidelberg, 211–223, 2008.
- [3] Yu, Z. Q.; Lu, L.; Guo, Y. W.; Fan, R. F.; Liu, M. M.; Wang, W. P. Content-aware photo collage using circle packing. *IEEE Transactions on Visualization and Computer Graphics* Vol. 20, No. 2, 182–195, 2014.
- [4] Bianco, S.; Ciocca, G. User preferences modeling and learning for pleasing photo collage generation. *ACM Transactions on Multimedia Computing, Communications, and Applications* Vol. 12, No. 1, Article No. 6, 2015.
- [5] Han, X. T.; Zhang, C. Y.; Lin, W. Y.; Xu, M. L.; Sheng, B.; Mei, T. Tree-based visualization and optimization for image collection. *IEEE Transactions on Cybernetics* Vol. 46, No. 6, 1286–1300, 2016.
- [6] Lee, M. H.; Singhal, N.; Cho, S.; Park, I. K. Mobile photo collage. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops, 24–30, 2010.
- [7] Mei, T.; Yang, B.; Yang, S. Q.; Hua, X. S. Video collage: Presenting a video sequence using a single image. *The Visual Computer* Vol. 25, No. 1, 39–51, 2009.
- [8] Liu, L.; Zhang, H.; Jing, G.; Guo, Y.; Chen, Z.; Wang, W. Correlation-preserving photo collage. *IEEE Transactions on Visualization and Computer Graphics* Vol. 24, No. 6, 1956–1968, 2018.
- [9] Kim, J.; Pellacini, F. Jigsaw image mosaics. *ACM Transactions on Graphics* Vol. 21, No. 3, 657–664, 2002.
- [10] Lee, H. Y. Automatic photomosaic algorithm through adaptive tiling and block matching. *Multimedia Tools & Applications* Vol. 76, No. 22, 24281–24297, 2017.
- [11] Pavić D.; Ceumern, U.; Kobbelt, L. GIZMOs: Genuine image mosaics with adaptive tiling. *Computer Graphics Forum* Vol. 28, No. 8, 2244–2254, 2009.
- [12] Goferman, S.; Tal, A.; Zelnik-Manor, L. Puzzle-like collage. *Computer Graphics Forum* Vol. 29, 459–468, 2010.
- [13] Huang, H.; Zhang, L.; Zhang, H. C. Arcimboldo-like collage using Internet images. In: Proceedings of the SIGGRAPH Asia Conference, Article No. 155, 2011.

- [14] Nan, L. L.; Sharf, A.; Xie, K.; Wong, T. T.; Deussen, O.; Cohen-Or, D.; Chen, B. Conjoining Gestalt rules for abstraction of architectural drawings. *Journal of Computer-Aided Design & Computer Graphics* Vol. 30, No. 6, 1–10, 2012.
- [15] Qi, Y. G.; Guo, J.; Li, Y.; Zhang, H. G.; Xiang, T.; Song, Y. Z. Sketching by perceptual grouping. In: *Proceedings of the IEEE International Conference on Image Processing*, 270–274, 2013.
- [16] Qi, Y. G.; Song, Y. Z.; Xiang, T.; Zhang, H. G.; Hospedales, T.; Li, Y.; Guo, J. Making better use of edges via perceptual grouping. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1856–1865, 2015.
- [17] Zhang, J. H.; Chen, Y. L.; Li, L.; Fu, H. B.; Tai, C. L. Context-based sketch classification. In: *Proceedings of the Joint Symposium on Computational Aesthetics and Sketch-Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering*, Article No. 3, 2018.
- [18] Lun, Z. L.; Zou, C. Q.; Huang, H. B.; Kalogerakis, E.; Tan, P.; Cani, M. P.; Zhang, H. Learning to group discrete graphical patterns. *ACM Transactions on Graphics* Vol. 36, No. 6, Article No. 225, 2017.
- [19] Liu, C.; Rosales, E.; Sheffer, A. StrokeAggregator: Consolidating raw sketches into artist-intended curve drawings. *ACM Transaction on Graphics* Vol. 37, No. 4, Article No. 97, 2018.
- [20] Judd, T.; Ehinger, K.; Durand, F.; Torralba, A. Learning to predict where humans look. In: *Proceedings of the IEEE 12th International Conference on Computer Vision*, 2106–2113, 2009.
- [21] Bylinskii, Z.; Judd, T.; Borji, A.; Itti, L.; Durand, F.; Oliva, A.; Torralba, A. Mit saliency benchmark. 2015. Available at <http://saliency.mit.edu/>.
- [22] Patterson, G.; Xu, C.; Su, H.; Hays, J. The SUN attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision* Vol. 108, Nos. 1–2, 59–81, 2014.
- [23] Cheung, V. Shape collage. 2007. Available at <http://www.shapecollage.com/>.

Zuyi Yang received his bachelor degree in the Artificial Intelligence Department, Xiamen University, China, in 2018. Currently he is a master candidate in Xiamen University. His research interests include computer graphics and layout optimization.

Qinghui Dai received his master degree in the Artificial Intelligence Department, Xiamen University, in 2018. Currently he is a technology staff member of Tencent. His research interest is in computer graphics.

Junsong Zhang received his Ph.D. degree in computer science from the State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China, in 2008. He is currently an associate professor at the Mind, Art and Computation Group, Artificial Intelligence Department, Xiamen University. His main research interests include computer graphics, human computer interaction, and cognitive science.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Other papers from this open access journal are available free of charge from <http://www.springer.com/journal/41095>. To submit a manuscript, please go to <https://www.editorialmanager.com/cvmj>.