Beyond Saliency: Assessing Visual Balance with High-level Cues

Baris Kandemir, Zihan Zhou, Jia Li, James Z. Wang*
The Pennsylvania State University, University Park, Pennsylvania, USA

ABSTRACT

Automatic composition optimization is a vital technique for computational photography systems. *Balance in composition* is one of the agreed-upon principles of aesthetics and is commonly employed as a visual feature in many computational aesthetics studies. It refers to an equilibrium of visual weights within composition. Existing composition optimization and aesthetic quality assessment systems utilize the saliency map to represent balance. However, saliency map methods fail to account for high-level visual features that are important for compositional balance. Our work establishes a framework for the purpose of evaluating the relationship between visual features and compositional balance. This provides a better understanding of compositional balance and help improve composition optimization performance. A dataset based on a human subject study was created with photos representing main balance concepts such as symmetric, dynamic balance, and imbalance. We take the visual center given by human subjects as the dependent variable and the center-of-mass for each type of visual features as the predictor variable. Based on a linear regression model, we can assess how much each type of visual features contributes to the prediction of the visual center. Our findings show that high-level visual elements can help increase prediction accuracy with significance on top of saliency maps. Specifically, extra information provided through human and dominant vanishing point detection is statistically significant for assessing balance in the composition.

KEYWORDS

Art, Computational Aesthetics, Composition, Saliency, Compositional Balance

1 INTRODUCTION

Capturing beauty has served as a longstanding quest in the history of humanity. This quest has led to efforts to understand creativity and the production process of aesthetic artifacts. Such efforts can be observed in the daily lives of individuals from professional artists to amateurs. For instance, hikers spend extra time taking pictures of scenery they like. A designer attempts to come up with an aesthetically appealing and attention-grabbing design for a Web site or even a t-shirt. Therefore, there is a need of clear-cut rules to construct good compositions in a controlled manner.

Figure 1: Balance in composition is an important aesthetic concept. It can be employed in different ways to elicit different visual impact on the viewers. In the first row, the first image shows an example of symmetric balance, whereas the second image is imbalanced. In the second row, we show two examples of dynamic balance.

To develop these rules and understand their mechanisms, psychology, psychophysics, and art experts have been conducting studies with human subjects [2, 23]. Fine arts experts have compiled aesthetic guidelines drawn from careful observation of aesthetic works and recognition of patterns among them [18]. Computer science has joined this quest, seeing it as an interesting application of artificial intelligence. The question "How can we make computers appreciate aesthetics?" has garnered significant attention. Data-driven approaches have improved a computer's ability to provide numerical values for images' aesthetic quality [5].

One of the aforementioned guidelines is balance in composition. A balanced composition can be described as a composition whose visual elements achieve visual equilibrium as a whole, whether it is a photograph or a painting. These elements could be objects, texture, color, and shapes, among others. In its simplest form, balance can be achieved by using pure symmetry along vertical or horizontal axes of the image (Fig. 1, first image). Meanwhile, an imbalanced image has the effect of creating tension and raising uneasy, disquieting responses in the viewer. For example, the purposefully placed yellow truck in the second image in Fig. 1 draws our attention to the lower-left corner of the image. However, there is another type of balance that is much more intriguing. The experts call it "dynamic balance" or "asymmetrical balance". This type of balance occurs when an area of a certain visual element is balanced

*B. Kandemir, Z. Zhou and J. Z. Wang are with the College of Information Sciences and Technology. J. Li is with the Department of Statistics. Emails: {bzk142,zuz22,jol2,jwang}@psu.edu .

ThematicWorkshops'17, Mountain View, CA, USA

DOI: https://doi.org/10.1145/3126686.3126712

out with another element with different visual properties [18]. The tree stump and milky way, or the tower and series of poles demonstrate this phenomenon. Balance is a highly subjective concept that different observers of the same photograph can have very different views. The subjectivity makes modeling it highly challenging.

The concept of balance plays an important role in computational aesthetic studies on the automated aesthetic layout of magazines and Web pages; automated image cropping and retargeting; seam carving; and aesthetic quality assessment [3, 13]. The representation of balance has been primarily calculated through the center of mass of a saliency map [20]. A well-studied topic in computer vision, the saliency map of an image indicates those points at which humans give more visual attention to the image. There have been different methods proposed to predict the saliency map of an image.

It seems natural to employ saliency maps to represent balance in composition, as saliency maps and balance are both related to visual attention. In this paper, we challenge this assumed relationship and investigate whether there are other elements that can improve the representation of visual balance for computational aesthetic systems. This relationship is bridged through an analysis of visual center and saliency maps. To the best of our knowledge, no study to date has created a larger dataset compared to empirical arts studies, where the relationship between aesthetics, visual balance, and visual elements have been investigated through subject studies with small image sets containing few hundred images. Therefore, we created an image dataset by compiling images from a popular photograph-sharing Web site. This dataset demonstrates different balance characteristics, according to art literature. The visual centers of the photos were obtained from individuals' responses to an online survey designed in line with empirical art studies [22]. Linear models were employed to measure the predictive power of saliency. The findings indicate that the predictive power of the linear model improves if the system accounts for the visual weight of humans, and dominant vanishing point (perspective point).

The research question guiding this study can thus be defined as "*How well can saliency maps predict visual balance? How can this prediction be improved?*" Our **contributions** are as follows:

- We created an image dataset that shows different balance characteristics, which are symmetry, asymmetrical (dynamic) balance, and imbalance, with visual center votes from people.
- We expanded on visual features other than saliency that may be related to visual balance.
- We evaluated the relationship between saliency maps and visual center, hence visual balance, and showed that including high-level features such as human and vanishing point detection improved performance.

2 RELATED WORK

Study of balance has been investigated from two aspects. First, art and psychology literature looked into how compositional balance works, and then, computer science community leveraged it as a component in a computational aesthetic framework. This section gives the related work to the mentioned efforts.

Balance in Art. Gestalt school, one of the most important schools in art, defines balance as *the equilibrium of visual weights in the design or artifact*, and it was considered as an important principle for aesthetics [8]. According to a discussion in [2], there was a perceptional field of push and pull in the frame, as a force field. The eye assigns an equilibrium position to an object within the push and pull field. If the object is not on one of these equilibrium points, the composition feels restless to the eye. Hence, balance is claimed to be one of the factors of aesthetics.

Balance phenomenon has been studied by empirical art researchers. A study where the participants were asked to put a fulcrum under "paintings of accepted merit", with original and cropped versions, was conducted [22]. It turned out most of the paintings were not balanced in the center of the frame. Another study that compared center of mass (CoM) of random and artistic photographs showed that CoM of artistic photos is aligned with axial center [23]. The same study demonstrated that the CoM can be shifted by cropping an image, and people prefer cropped pictures where CoM aligned with the axial center. However, these studies failed to cover a larger picture base to test these findings and the pictures used were less related to multimedia.

Balance in Multimedia and Vision. Compositional balance has raised interest in multimedia and computer vision communities as computational aesthetics became an interesting research topic. The subjectivity of aesthetics has been widely accepted by different research communities; however, data-driven approaches have shown that some consensus can be reached [15]. Upon this, the data-driven approach found another area of application to make visual data such as photographs automatically more aesthetic.

In computational aesthetics, the balance has been considered as a small feature to be included with some ad hoc methods and its definition has not been clear. It was employed in composition optimization, image attention retargeting, seam carving, automated layout design, and photo quality assessment systems [10, 13, 20, 21, 26, 36]. The bottom-up saliency maps of images were utilized to quantify balance in compositional optimization [20]. In photo aesthetic quality assessment, saliency is utilized to measure how much a photo adheres to balance and the rule of thirds principles [21]. Bottom-up saliency map was employed to place text on less busy parts of the background image for automated magazine cover design [13]. These approaches were based on the center of mass and rely on low-level features, *e.g.*, local orientations and intensity. However, they did not take other factors into account such as object or shapes whereas balance is related to higher-level features [17].

3 MOTIVATION

In this section, we expand on saliency map concept that is frequently employed by aforementioned computational aesthetic studies, and state the problem by presenting cases where bottom-up saliency map methods fail in visual balance representation.

Saliency. The salient regions of an image refer to the parts of the image on which the viewer's eyes spend more time (fixate) while seeing the image. In particular, "saliency intuitively characterizes some parts of a scene–which could be objects or regions–that appear to an observer to stand out relative to their neighboring parts" [4]. Saliency has been considered from two cognitive angles, which are bottom-up and top-down. Top-down approaches focus on tasks, memory, learning and related heuristics [27]. Saliency is

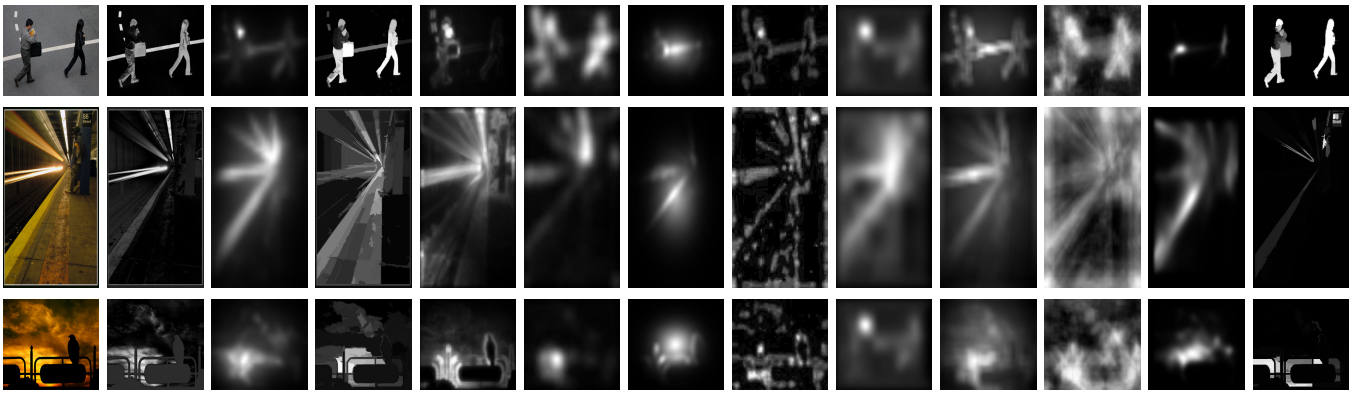

Figure 2: Saliency analyses often fail to capture the key elements in attention. Columns: original image, Achanta, GBVS, DRFI, Context aware, imSig, covSal, corSal, UHF, SWD, Murray, FES, MDF.

Category	Name	Description
Patch-based	DRFI [14]	Dissimilarity measure defined among segmented regions in multi scales.
	Context Aware [9]	Dissimilarity defined among patches with a spatial distance constraint for context.
	Achanta [1]	Frequency information through different channels is leveraged for dissimilarity.
	CovSal [7]	Covariance among patches is employed for dissimilarity information.
	SWD [6]	The dissimilarity patches represented in reduced dimensional space is utilized.
Graph-based	GBVS [11]	Pixel information such as color, intensity, orientations are put into a graph, where edges represent a dissimilarity measure. The graph is treated as a Markov Field to obtain an activation map.
Center-surround	CorSal [30]	The corner cues obtained via application of Gabor filters are used for saliency information.
	FES [33]	Bayesian framework between visual features and saliency within a moving center-surround window.
	Murray Model [25]	Center-surround filter size and other parameters are learned through GMM.
Others	ImSig [12]	Saliency map based on Inverse DCT of image signature for foreground-bacground separation.
	UHF [32]	Unsupervised hierarchy of visual features for saliency
	MDF [19]	Multi scale feature learning through CNN for saliency inference.
	Shallow Convnet [28]	A shallow neural network is followed by a deep network in order to regress saliency values.

Table 1: State-of-the-art bottom-up saliency methods are summarized according to their approaches.

more tailored according to the task at hand, which, for instance, may be counting humans in the picture.

Bottom-up models focus on low-level visual properties. This class of approaches attempts to fuse different low-level features in accordance with how human visual system works [35]. The theory states that primate visual system pays extra visual attention to different or anomalous regions in the image. The dissimilarity of regions within an image is represented via fusion of low-level visual features such as local orientations, texture, colors, curvatures, and intensity. Table 1 summarizes approaches to capture saliency which can be further categorized into (i) graphical models; (ii) patch-based models; (iii) cognitive models; and (iv) others .

Observations and Problem. As explained above, bottom-up approaches consider pre-attentive visual attention that plays a role in viewer's unconscious reaction. This nature constrains these methods more to low-level features which are processed at earlier stages of the human visual system. When the center of mass of a bottom-up saliency map is evaluated to check whether visual attention is lopsided or aligned with the physical center of the image, some problems arise. Fig. 2 shows a few examples where saliency methods fail to capture the elements that actually have more or less visual weight than saliency map assigns to them. The first picture in Fig. 2 contains two humans that balance each other visually. At first glance, almost half of the methods tested fail to capture humans, whereas it was shown in a study that humans, faces, body parts, text bodies, and animals get more attention than usual [16]. Among the methods tested, the ones that successfully capture humans in pictures assign attention to other parts of images which actually do not have that much visual weight compared to humans. The text area above the person in the second image is also mostly missed, but may not have that much visual weight as it covers a small area in the image. More interestingly, saliency maps employed capture the light beams of the subway and lines created by subway station floor to great extent. An interesting point about these lines is that they are parallel in reality, but appear to converge on the photograph plane. This structure leads the eye from left to the convergence point, which has a different visual weight. Saliency methods do not succeed in reflecting this concept. The third image shows another failure of saliency map regarding images that contain high contrast regions. The bird silhouette perched up on the railings is surrounded with a relatively light region of

clouds. Saliency methods successfully get the outline of the bird and the railing bars, which are regions of high contrast. But in a higher scale the bird itself is different from the cloud structure, so it has more visual weight. These imperfections in saliency maps regarding visual weights indicate that a new approach that takes high-level features into account can be highly beneficial.

4 THE METHOD

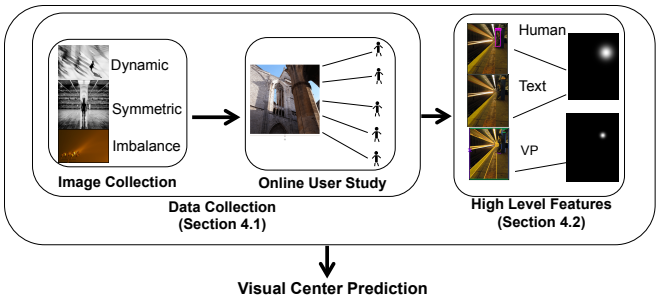

Figure 3: Our framework for visual center prediction.

Our visual center prediction framework contains four parts as summarized in (Fig. 3). First two parts are related to obtaining the image set and participant ratings related to visual center of the images. The third component is the extraction of low- and high-level features, and translating into a vector representation. Finally, we predict the visual center through linear regression scheme. This section provides details about the first three components.

4.1 Data Collection

We detail the image set creation, interface design for acquiring ground-truth labels, and feature extraction from the image set.

Image Set. The visual data for the initial study was mainly collected from a popular photo-sharing Website, Flickr, using the query 'urban' and they were ranked according to their interestingness. These images were categorized into four classes, 'symmetrical balance', 'dynamic balance', 'imbalanced', and 'hard to tell', by the authors based on consensus. While categorizing the images, the definitions in [17, 18] were taken as merit. After eliminating images that are not related to balance concept, 779 images were collected in total. In the end, the number of samples in these classes are 41, 362, 90, and 447, respectively. The fact that the total number of balanced images was 403 and that for imbalanced images was 90 caused an imbalance in the data. Hence, 286 more imbalanced images were selected based on consensus from the query results for 'travel', 'Hawaii', 'jungle', 'party', 'galaxy', and 'vacation'. Hence we had 403 balanced and 376 imbalanced images. The reason these query words were selected is that urban pictures were dominated by man-made structures. By adding images belonging to these query we ensured that more natural scenic images with different properties were included in our dataset.

Interface Design. An online study was designed to collect data from the participants. The challenges of designing an online survey about compositional balance started with quantification or measure of balance. In empirical art field, some studies put a fulcrum under the picture and asked participants to move the lever till the picture feels balanced on the fulcrum [22, 23]. In these studies, the normalized balance scores were recorded in an interval [-1, 1], where 0 was the exact center of an image, and -1 and 1 are the extreme edges. In a similar fashion, this concept was transferred to our online study, this is one of the firsts in online balance studies to the best of our knowledge. We recorded the balance values in the interval [0,100], where 50 was the exact center of the image and 0 and 100 represented the extreme edges (Fig. 4).

Figure 4: User interface for online human subject study. The slider bar below the image allows the participant to indicate the visual center.

The second challenge was coming up with a concise meaningful tutorial that included a definition of compositional balance and how it was achieved. As mentioned in the previous section, simple definitions of compositional balance and its methods were given as in [17, 18] along with sample images. Once the participant selects a participant ID, he/she must read through a tutorial explaining all of the concepts that are being studied in the survey. The ways compositional balance is achieved are detailed with example images and text in the tutorial which also set the same rules for the compilation of the image set. After the tutorial, the participant enters the survey. The survey contains a white background with the test image in the center-left, a slider bar beneath the image. The slider bar starts with the mark on the center - which correlates to 50 in our study. The participant is asked to 'move slider in a way that the image feels visually balanced at that point'. If the participant needs any assistance or wants to refer to the tutorial, he/she is free to by going back "Home". For each image, we limited the number of ratings to five. Each participant is only presented with a specific image once. In addition, to encourage breaks, each participant only does a batch of 100 images at once. When the participant has completed ratings for 100 batches, or there are no more images to rate, he/she will enter a screen informing him/her of that. It's also possible to start a new batch right after one is completed. If the participant has been inactive on one single page for up to six minutes, the survey

will time out and the participant will be returned to the home page. By the end of the survey time period, the important data we were able to collect consisted of the participant IDs, ratings of the center of mass. The study was taken by eight participants.

Human Subject Study and Data Properties. The ratings were collected from undergraduate, graduate students and faculty who participated in the study. Each image received five ratings. The average rating position of slider, $\bar{p}_i$, for each image i is calculated. The mean slider position across all images p^* is 50.17 with standard deviation of 6.7.

$$\bar{p}_i = \frac{1}{N} \sum_{u \in \text{Participants}_i} p_i^u \ , \quad p^* = \frac{1}{M} \sum_i \bar{p}_i \ , \tag{1}$$

where Participants_i is the set of participants who have rated for the image i. When images with highest standard deviation in slider position were inspected, the border cases where people showed disagreement were observed. It was understood that people may have different opinions of symmetry, and were confused with dynamic balance for some cases. It was also observed that some participants agreed on the imbalanced category of an image, but the slider ratings pointed to different sides for imbalance. This seems to indicate that people can have a different understanding of the concept. These results showed that there were cases of agreement for images which raises hope to achieve a consensus.

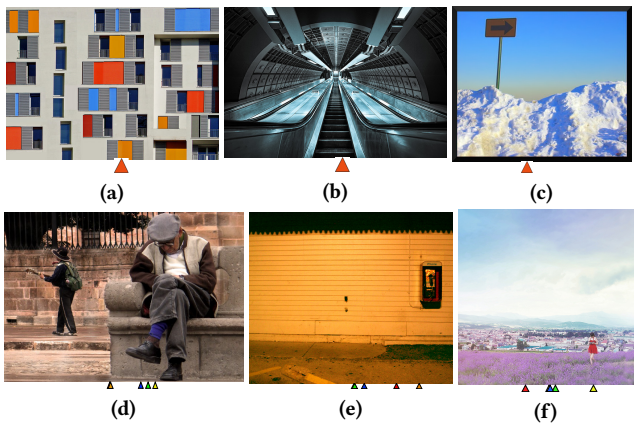

Figure 5: Example images with low and high standard deviation for average slider position. Images with lower standard deviation are shown in first row, where slider positions overlap and the position is illustrated with a single triangle. Second row shows images with higher standard deviation. Participant ratings are marked as triangles to show the spread of slider positions.

The images where participants had a high level of agreement were chosen through a standard deviation threshold on slider positions. The standard deviation of slider position rates is expected to be low where participants agree. The thresholding process resulted in a total of 593 images with a high level of agreement. Fig. 5 displays some pictures with low (first row) and high (second row) standard deviation. As the high standard deviation images are observed, it is seen that people assign different weights to different parts for the same picture. While some participants gave more weight to the sitting man closer to the camera or the girl in red dress, some deemed the weight distribution equal (Fig. 5d and 5f). The weight of the dark phone booth varies across participants as seen in Fig. 5e.

4.2 High- and Low-Level Features For Balance

This subsection explores the features that may be helpful in predicting the visual center of an image.

Saliency. As a model of visual attention distribution over an image, saliency maps were employed. Bottom-up saliency approaches suit more to our purpose, because we are interested in unconscious visual response to the composition balance, which depends on low-level features. A representative of each category of saliency methods in Table 1 were selected for our study, including context-aware saliency, image signature, graph-based, UHF, MDF and non-parametric low-level vision saliency (Murray).

Informative Objects. People pay more attention to the objects that provide different kinds of information about the picture, which can be considered as high-level features. The information can be emotional cues, actual knowledge, or message. One of the prominent source of information in pictures are humans. One another object class is text in the image. Inclusion of the spatial distribution of these object along with saliency could improve the representation accuracy of visual balance. Out of 779 images in our dataset, there were 228 images that contained humans. Apart from that, there were 161 images that contained text regions.

For human detection, state-of-the-art scene annotation system based on deep neural networks, YOLO, was utilized [29], where the model was trained on ImageNet 1000-class competition dataset [31]. In this approach, a single convolutional network is run on the whole image. The network returns bounding boxes of objects detected with a probabilistic confidence. Finally, the system thresholds the bounding boxes by their corresponding probabilities. The method employed was successful to return human positions for 226 out of 228 images. The recall of this method on our dataset was 0.99 in terms of detecting humans. For text detection, another deep learning based method, named 'connectionist text proposal method' proposed in [34] was employed. This approach detects text line by densely sliding a small window (3x3) in the convolutional feature maps and produces fine-scale text proposals. Fine-scale text proposals are susceptible to false detection for windows, bricks, etc. The sequential structure of text is exploited via final recurrent neural network to improve performance. The recall rate of text detection on our dataset was 0.89 . The detection boxes are eliminated by considering the aspect ratio, as really thin text or humans/human parts are normally not recognized by people.

Eye Leading Lines. As aforementioned, one of the elements that are used to strike compositional balance is eye leading lines (Fig. 6). The concept of eye leading lines in visual art or photography is related to the perspective, which is mainly associated with prominent geometrical structures called *parallel lines* [37]. On photograph plane, parallel lines appear to converge towards a point, which is called the vanishing point (VP). There may be more than one VPs in a scene, but the main interest is on the dominant one. In our approach, we leverage dominant vanishing point detection method

proposed in [37] to account for eye leading lines information, which can be seen as a high-level feature.

In our approach, straight edges are extracted from ultrametric contour maps by subdividing the contour into straight line segments at points that have maximum distance to the straight line connecting the end points of the contour. Then these straight edges are grouped according to J-Linkage fitting. Two random edges (E_{j1}, E_{j2}) are sampled from the edge set. Lines are fitted to each edge, their hypothetical intersection point v_j is obtained. J-Linkage creates a preference matrix according to a consistency measure. Consistency measure was defined as the root mean square of distance between edge points and a line, $\hat{l}$ that passes through the hypothetical v_j and minimizes this distance. Once the preference set for each edge is acquired, edges that have similar preference sets are clustered together. v_j of the biggest cluster becomes the dominant VP.

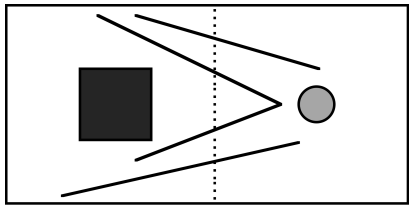

Figure 6: Eye leading lines are utilized to strike a balanced composition [17].

4.3 Representation

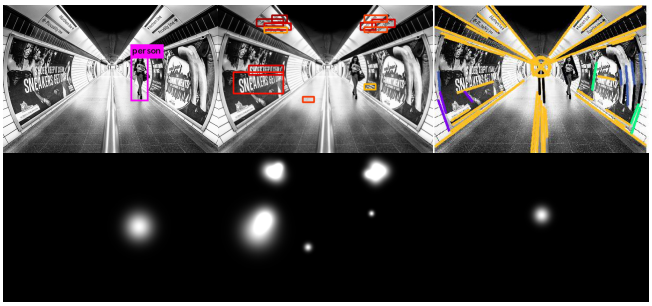

Figure 7: Representation of different features. Gaussians are fitted within the boxes obtained and CoM of this image is calculated. From top left to bottom right: YOLO human detection, text detection, dominant VP detection, and the corresponding Gaussian envelopes.

Current literature in psychology and computer science borrows the physical concept of center-of-mass (CoM) given in Eq. (2). In this case, the visual weight elements are modeled as point-masses. CoM coordinates give an idea about into which quadrant of the image the visual weight falls.

$$CoM = \frac{1}{M} \sum_{m_x \in X} m_x \times \vec{r}\,, \tag{2}$$

where m_x is a point mass in set X, M is total mass, and $\vec{r}$ is the position vector of m_x . As saliency maps were in grayscale, it was relatively easy to compute CoM. However, there was a question of how to calculate CoM for human-text, and vanishing point detection. In [11], visual attention on a region is enveloped with a Gaussian distribution, as the most attention is paid to the center and it degrades further from the center. We adopt this concept and fit a Gaussian within human, text detection boxes and around the vanishing point detected (Fig. 7). Very thin boxes with high aspect ratio were eliminated. The upper left corner coordinates, width and height of each box in the image is taken and a Gaussian is fit inside the box for the remaining boxes (Eq. (3) and (4)). The spread of the Gaussian becomes a parameter to be tuned according to each box. The value σ is adjusted according to the diagonal length of the rectangle. The Gaussian that is fitted to i-th box b_i is

$$\sigma = \sqrt{\text{height}_{b_i}^2 + \text{width}_{b_i}^2}\,, \tag{3}$$

$$G = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x - x_{CoM})^2 + (y - y_{CoM})^2}{2\sigma^2}\right\}\,. \tag{4}$$

Another concept that can be borrowed from physics is representing forces as vectors. The coordinates of CoM can be transformed to a force vector. The vectors computed from CoM of different image information channels can represent visual forces caused by different visual sources, which models the push and pull. The coordinates of CoM is converted into a vector where the origin is the center of the image. Our approach is to combine CoM vectors through a vectorial sum. The values of vectors are standardized between 0 and 100.

5 EVALUATION

After representing each visual weight component in a vectorial force structure, evaluation of how well these forces predict visual center of an image in relation to visual balance follows. This section describes the evaluation setup and provides the results.

Setup. In line with the competing visual forces mentioned in [2], we obtained the CoM vectors for saliency, human, text and VP detection. However, the way these visual forces interact were unknown to us. As first criteria for performance, we inspected the absolute value of the difference between participant visual center values and CoM calculated for different features, $|\textbf{center}_{\text{participant}} - \textbf{center}_{\text{feature}}|$. A sorted version of these differences were evaluated and checked whether the feature in question causes a drop in difference measure. It was observed that a simple vectorial sum failed to capture the actual resultant vector as the contribution of each element was not known with certainty.

The problem can be addressed through a linear regression system as explained in the following. The performance measure utilized for prediction is mean square error in a cross validation setting.

Contribution of each feature can be explored through a linear model that combines the forces. The problem can be considered as a regression problem where the dependent variable is the location of visual center and predictors are vectors for the visual forces at play. In order to show the effect of each element for prediction, we devised a hierarchical regression method that started with a base model which only had saliency information and adds each other

feature gradually:

$$\begin{aligned}
\text{(Model 1) visual_center} =& \beta_0 + \beta_1 \cdot \text{CoM}_{\text{Saliency}} \ , \\
\text{(Model 2) visual_center} =& \beta_0 + \beta_1 \cdot \text{CoM}_{\text{Saliency}} + \beta_2 \cdot \text{CoM}_{\text{Human}}, \\
\text{(Model 3) visual_center} =& \beta_0 + \beta_1 \cdot \text{CoM}_{\text{Saliency}} + \beta_2 \cdot \text{CoM}_{\text{Human}} \\
& + \beta_3 \cdot \text{CoM}_{\text{Text}} \ , \\
\text{(Model 4) visual_center} =& \beta_0 + \beta_1 \cdot \text{CoM}_{\text{Saliency}} + \beta_2 \cdot \text{CoM}_{\text{Human}} \\
& + \beta_3 \cdot \text{CoM}_{\text{Text}} + \beta_4 \cdot \text{CoM}_{\text{VP}} \ .
\end{aligned}$$

The hierarchical analysis showed that addition of human and VP detection information demonstrated statistically significant ($p < 0.05$) improvements to the base model while text detection information had no significant effect. The models were also tested with 3-fold cross validation and performance of models were measured through mean square error (MSE) for each model. The results are shown in Fig. 8.

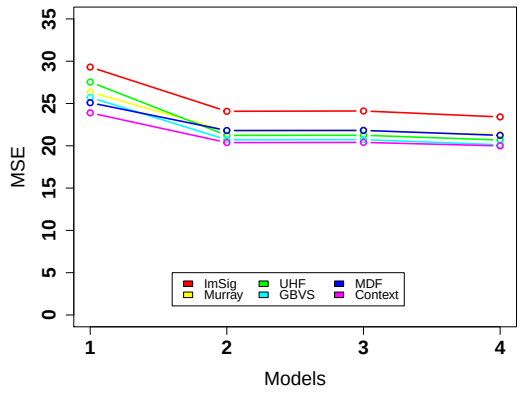

Figure 8: MSE values of model 1, model 2, model 3, and model 4 for selected saliency algorithms. Each line corresponds to a saliency algorithm chosen to represent categories given in Table 1.

Results. The analyses showed that saliency was useful for representation of compositional balance, but was not enough. Further analysis demonstrates that there is room for improvement. In light of empirical art studies, high-level features that may enhance saliency map performance were explored. The MSE analysis showed that the addition of human information contributed to the representation power as seen in first two images in the first row of Fig. 9. This contribution can be based on two observations. First, the undetected humans by saliency maps are sensed through including human detection information. Second, consolidating information regarding human positions helps suppress noisy saliency components.

Third to fifth images in the first row of Fig. 9 portray why the dominant VP detection information helped boosting prediction performance. Saliency maps fails to account for the effect of parallel long lines leading the eye toward one side of the picture, *i.e.* ceiling lines, station floor, or the building edges for the last image. The saliency values pick upon the components that make up parallel lines, which is mainly what bottom-up models are designed to achieve, but it fails to capture their influence on balance. By including this information, there is improvement in the prediction performance.

On the other hand, including the text detection information did not provide any improvement to the base model. When the cases that include text are considered, this situation can be interpreted in two ways. In the first case, the text information was not as important as we had expected, so saliency methods failing to detect text area did not negatively affect the performance. The other case may be that the contrast difference around the text area and high-frequency texture of text was already captured via the bottom-up saliency method and adding text location did not do anything more than stating the obvious.

Some cases where our model did not improve prediction based solely on saliency or aggravated it are also shown in Fig. 9 (lower row). Basically these are the cases where there are no dominant human, or vanishing point components. The traffic lights in the first image in that row is the main object in the image which is captured by saliency. As there is no other components that our model includes, it predicts average response based solely on saliency information, which makes it worse. In the second image, the outline of the frog is captured by the saliency which is in line with participant ratings. Average response from our model makes it worse as the saliency CoM is multiplied with a coefficient. The third and fourth images have objects of different colors, *i.e.* red tree and walls of different colors. The saliency map can capture the tree based on color edges, while the model exacerbated the prediction due to regression coefficients. The contrast component, the fire, of the last image gets more attention than the humans walking away from the car, which is not sensed by our model.

6 DISCUSSION

The study suggests that there are high-level visual elements that are influential in compositional balance but are not captured by bottom-up saliency maps. Accounting for these elements decreased the error compared to visual center prediction employing only saliency maps. Inspection of the failure cases shows that there is room for further improvement. More compositional components for further consideration are explored during our study. These components can be color and contrast. As seen in the lower row of Fig. 9, the color of the tree and the color of walls create different visual weights. Saliency captures the visual center correctly as it senses the edge structures of the tree and the windows. However, the color information is not encoded. One further step could be accounting for different visual weight of different color areas [24]. The last image in the row suggests objects of contrast may have different visual weight.

In [16], saliency maps were learned from actual gaze maps, and it was concluded that humans, body parts, animals and text get more attention in an image. As a top-down approach was adopted in this study, the contribution of each visual element was not studied. With our study, we attempted to analyze each element.

The hierarchical linear regression utilized in our study may have done a good job in testing for contribution of added high-level

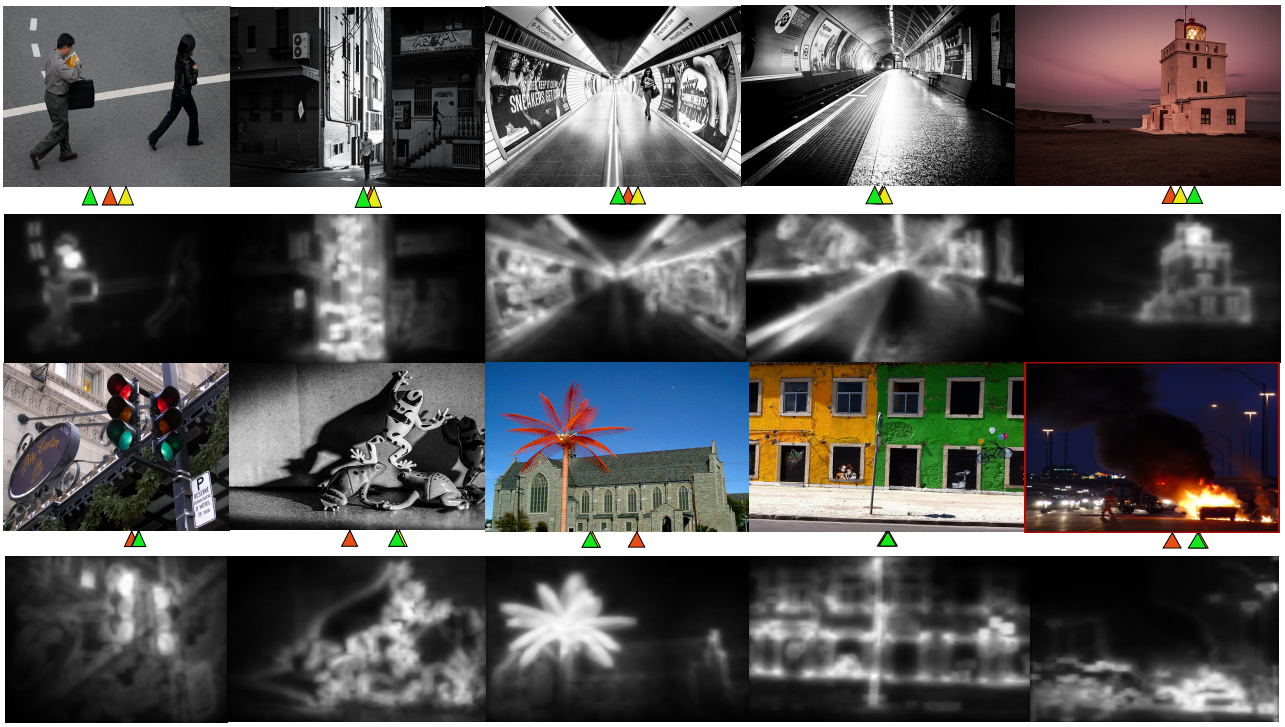

Figure 9: Evaluation on visual center prediction. The yellow triangle shows mean visual center position annotated by participants. The green triangle is the CoM of saliency images. The red triangle is the prediction of the proposed model. The top row demonstrates cases where our model improved prediction along with the saliency maps (context aware saliency). The lower row displays cases where the model was worse than saliency or didn't improve prediction.

features. One aspect that may not be captured via simple linear regressions is interaction between these features. Furthermore, they may also fail to detect nonlinear relationship between the features and visual center locations given by participants. More complex learning schemes can be employed.

Another aspect of the study that can be improved is the size of the image data. The number of images can be considered large compared to empirical art studies, but it can be larger to be more comprehensive. As there are no previous dataset regarding this problem, the pictures had to be selected carefully to certainly show the balance characteristics. This process required inspection of pictures manually, which constrained the number of images substantially. Another point about dataset collection is the online survey. As the study was conducted with a group of graduate students in a controlled lab environment, we deemed detection of uninformative annotators unnecessary. If a much larger participant base is to be used, elimination of uninformative annotators needs be incorporated. In our study, the participants were from a small pool of graduate students whose demographics were not recorded. For next stage of the visual balance research, the influence of demographics can be further explored. In terms of different balance characteristics, next step is the association of visual center to dynamic balance and imbalance.

For future, the implementation of other contributing high-level features and analysis of their effect on prediction performance may be helpful to pinpoint visual elements in the composition that can be used to strike a balanced structure. As the visual center prediction improves, the seam carving, image thumb-nailing, and retargeting applications may perform better.

7 CONCLUSIONS

We investigated the relationship between compositional balance and saliency concept through saliency maps' predictive power for visual center. In order to achieve this, we compiled a dataset of images that are rated by participants through an online survey. Different bottom-up saliency methods were run on these images. Center of mass of saliency maps were directly utilized to predict visual center. On top of saliency information, we included human, text and vanishing point detection information based on theories in empirical art studies. The analyses demonstrated the addition of human and vanishing point detection on top of saliency improved prediction of the location of visual center of images.

Acknowledgments: This material is based upon work supported in part by the National Science Foundation under Grant No. 1110970. Edward Chen assisted in the data collection and data analysis effort.

REFERENCES

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. 2009. Frequency-tuned salient region detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 1597–1604.

[2] Rudolf Arnheim. 1954. *Art and Visual Perception: A Psychology of the Creative Eye*. Univ. of California Press.

[3] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. 2010. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of International Conference on Multimedia*. ACM, 271–280.

[4] Ali Borji and Laurent Itti. 2013. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2013), 185–207.

[5] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2006. Studying aesthetics in photographic images using a computational approach. *Proceedings of the European Conference on Computer Vision* (2006), 288–301.

[6] Lijuan Duan, Chunpeng Wu, Jun Miao, Laiyun Qing, and Yu Fu. 2011. Visual saliency detection by spatially weighted dissimilarity. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 473–480.

[7] Erkut Erdem and Aykut Erdem. 2013. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision* 13, 4 (2013), 11–11.

[8] Sharon Gershoni and Shaul Hochstein. 2011. Measuring pictorial balance perception at first glance using Japanese calligraphy. *i-Perception* 2, 6 (2011), 508–527.

[9] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. 2012. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 10 (2012), 1915–1926.

[10] Y.W. Guo, M. Liu, T.T. Gu, and W.P. Wang. 2012. Improving photo composition elegantly: Considering image similarity during composition optimization. In *Proceedings of Computer Graphics Forum*, Vol. 31. Wiley Online Library, 2193–2202.

[11] Jonathan Harel, Christof Koch, and Pietro Perona. 2007. Graph-based visual saliency. In *Advances in neural information processing systems*. 545–552.

[12] Xiaodi Hou, Jonathan Harel, and Christof Koch. 2012. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 1 (2012), 194–201.

[13] Ali Jahanian, Jerry Liu, Qian Lin, Daniel Tretter, Eamonn O'Brien-Strain, Seungyon Claire Lee, Nic Lyons, and Jan Allebach. 2013. Recommendation system for automatic design of magazine covers. In *Proceedings of International Conference on Intelligent User Interfaces*. ACM, 95–106.

[14] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. 2013. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 2083–2090.

[15] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. 2011. Aesthetics and emotions in images. *IEEE Signal Processing Magazine* 28, 5 (2011), 94–115.

[16] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *Proceedings of International Conference on Computer Vision*. IEEE, 2106–2113.

[17] John Kahrs, Sharon Calahan, Dave Carson, and Stephen Poster. 1996. Pixel cinematography: a lighting approach for computer graphics. *ACM SIGGRAPH Course Notes* (1996), 433–42.

[18] David A. Lauer and Stephen Pentak. 2011. *Design Basics*. Cengage Learning.

[19] Guanbin Li and Yizhou Yu. 2015. Visual saliency based on multiscale deep features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 5455–5463.

[20] Ligang Liu, Renjie Chen, Lior Wolf, and Daniel Cohen-Or. 2010. Optimizing photo composition. In *Proceedings of Computer Graphics Forum*, Vol. 29. Wiley Online Library, 469–478.

[21] Wei Luo, Xiaogang Wang, and Xiaoou Tang. 2011. Content-based photo quality assessment. In *Proceedings of International Conference on Computer Vision*. IEEE, 2206–2213.

[22] I.C. McManus, D. Edmondson, and J. Rodger. 1985. Balance in pictures. *British Journal of Psychology* 76, 3 (1985), 311–324.

[23] I.C. McManus, Katharina Stöver, and Do Kim. 2011. Arnheim's Gestalt theory of visual balance: Examining the compositional structure of art photographs and abstract images. *i-Perception* 2, 6 (2011), 615–647.

[24] Robert H. Morriss and William P. Dunlap. 1988. Influence of chroma and hue on spatial balance of color pairs. *Color Research & Application* 13, 6 (1988), 385–388.

[25] Naila Murray, Maria Vanrell, Xavier Otazu, and C. Alejandro Parraga. 2011. Saliency estimation using a non-parametric low-level vision model. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 433–440.

[26] Tam V. Nguyen, Bingbing Ni, Hairong Liu, Wei Xia, Jiebo Luo, Mohan Kankanhalli, and Shuicheng Yan. 2013. Image re-attentionizing. *IEEE Transactions on Multimedia* 15, 8 (2013), 1910–1919.

[27] Aude Oliva, Antonio Torralba, Monica S. Castelhano, and John M. Henderson. 2003. Top-down control of visual attention in object detection. In *Proceedings of The International Conference on Image Processing*, Vol. 1. IEEE, I–253.

[28] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E. O'Connor. 2016. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 598–606.

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 779–788.

[30] Wirawit Rueopas, Sangsan Leelhapantu, and Thanarat H. Chalidabhongse. 2016. A corner-based saliency model. In *Proceedings of International Joint Conference on Computer Science and Software Engineering*. IEEE, 1–6.

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[32] Hamed R. Tavakoli and Jorma Laaksonen. 2016. Bottom-Up Fixation Prediction Using Unsupervised Hierarchical Models. In *Proceedings of Asian Conference on Computer Vision*. Springer, 287–302.

[33] Hamed R. Tavakoli, Esa Rahtu, and Janne Heikkilä. 2011. Fast and efficient saliency detection using sparse sampling and kernel density estimation. In *Proceedings of Scandinavian Conference on Image Analysis*. Springer, 666–675.

[34] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. 2016. Detecting text in natural image with connectionist text proposal network. In *Proceedings of the European Conference on Computer Vision*. Springer, 56–72.

[35] Anne Treisman. 1985. Preattentive processing in vision. *Computer vision, graphics, and image processing* 31, 2 (1985), 156–177.

[36] Xianjun Sam Zheng, Ishani Chakraborty, James Jeng-Weei Lin, and Robert Rauschenberger. 2009. Correlating low-level image statistics with users-rapid aesthetic and affective judgments of web pages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1–10.

[37] Zihan Zhou, Farshid Farhat, and James Z. Wang. 2017. Detecting dominant vanishing points In natural scenes with application to composition-sensitive image retrieval. *IEEE Transactions on Multimedia* (2017).