Semi-supervised Learning Based Aesthetic Classifier for Short Animations Embedded in Web Pages

Dipak Bansal, Samit Bhattacharya

Semi-Supervised Learning based Aesthetic Classifier for Short Animations Embedded in Web Pages

Dipak Bansal[1] and Samit Bhattacharya[1]

[1] Dept. Of Computer Science and Engineering, IIT Guwahati-781039, Assam, India

{d.bansal143, samit3k)@gmail.com

Abstract. We propose a semi-supervised learning based computational model for aesthetic classification of short animation videos, which are nowadays part of many web pages. The proposed model is expected to be useful in developing an overall aesthetic model of web pages, leading to better evaluation of web page usability. We identified two feature sets describing aesthetics of an animated video. Based on the feature sets, we developed a Naïve-Bayes classifier by applying Co-training, a semi-supervised machine learning technique. The model classifies the videos as *good*, *average* or *bad* in terms of their aesthetic quality. We designed 18 videos and got those rated by 17 participants for use as the initial training set. Another set of 24 videos were designed and labeled using Co-training. We conducted an empirical study with 16 videos and 23 participants to ascertain the efficacy of the proposed model. The study results show 75% model accuracy.

Keywords: Aesthetics, web page, short video, classification, semi-supervised learning, Co-training.

1 Introduction

Usability professionals over the years have been working extensively on developing methods and techniques to determine the usefulness of interactive systems. These activities are sought to be augmented in recent years with the studies on measuring perceived usability of the system, which relies heavily on aesthetics [7]. Postrel [9] contented that the 21st century is the "age of aesthetics". The contention may well be true in the context of interactive systems, as the large number of recent works show [3,17,20,28-30].

Web pages are good examples to consider the importance of aesthetics in interactive system design. Most of the pages contain various types of information, put together using various design patterns. Consequently, the complexity of the interfaces in terms of information content and layout is usually high. Evidently, the aesthetics of the design determines to a great extent its acceptability (and therefore, usability) to the users [11, 12, 18, 19, 22, 23].

While the role of aesthetics on usability is clear, the problem lies in measuring it. Usability studies depend on quantifiable measurements. However, development of such measures for evaluating aesthetic quality of an interactive system is still in its infancy, primarily because of the perception that "aesthetics is subjective." While it is true to an extent, it is not impossible to develop quantitative measures of web page aesthetics, as reported in [16, 24]. Both these works actually demonstrate the possibility of computational modeling of aesthetics. The advantage of having a computational model is in the ability to evaluate aesthetic quality of an interface *automatically*, thereby making it possible to integrate the model as a tool in a web page design environment so that the designer can check their design quickly.

In this work, we propose a model to compute aesthetics of short animation videos, which are embedded in many of the web pages nowadays. This work is part of a larger goal of computational modeling of whole web page aesthetics. We base our work on the philosophy that modeling component aesthetics and then combining those models will lead to an overall web page aesthetics model.

We propose a semi-supervised learning model to compute aesthetics of short videos. The model is essentially a Naïve-Bayes classifier, which classifies a video into one of the three classes: *good*, *average* and *bad* with respect to its aesthetic quality. On the basis of a study of 18 web pages and prior work on this field, we identified two feature sets to capture short video characteristics with respect to its aesthetic quality. The feature sets are used in a co-training method to develop the classifier. We validated our model with an empirical study involving 16 videos and 23 participants. The feature sets, the co-training method and the validation study are described in this paper.

The paper is organized as follows. In Section 2, we discuss the related works reported in the literature along with their limitations that we address in this work. The development of the proposed model is detailed in Section 3. The empirical study conducted to evaluate the model performance is reported in Section 4. In Section 5, we discussed the strengths and limitations of the proposed model along with the scope for further works. The paper is concluded in Section 6.

2 Related Works

Considered as a branch of philosophy that deals with the nature of beauty, art, and taste[1], aesthetic design has been extensively studied in the field of fine and commercial arts [1, 7]. The importance of aesthetics in human affairs has been elaborated by Maquet [13]. In fact, as early as 1984, the role of aesthetics in determining usability of interactive systems was highlighted [9], where it was reported that a poorly designed computer screen can hinder communication. The positive effect of good graphic design and attractive displays on the transfer of information has been found by Aspillaga [2]. Elements of aesthetic considerations were present in other works as well [21,31-33].

[1] From Wikipedia. See http://en.wikipedia.org/wiki/Aesthetics

Despite the presence of such early works, only the later part of the 1990s saw a spurt in activities in this area. These works included investigation of the role of aesthetics on interactive system design in general as well as on the effects of aesthetics in specific interaction domains. Researchers argued about the role of aesthetics in interactive system design [27]. Set of guidelines for screen design, keeping in mind the aesthetic aspect, were proposed [8]. In the context of e-learning, the effect of aesthetically pleasing layouts on the student's motivation to learn has been reported [26]. Szabo and Kanuka [25] found that subjects who used the lesson with *good design principles* completed the lesson in less time and had a higher completion rate than those who used the lesson with poor design principles.

A typical scenario where aesthetics play important role in the overall usability of the system is the design of web pages. Relationship between visual appeal and perceived usability of web pages was investigated in Lindgaard et al. [12]. Schmidt et al. [23] found correlation between usability and aesthetics in the context of subjective evaluation, depending on the user's background, goal, task, and application type. Several works concentrated on developing measures to assess aesthetic quality of web pages [11, 15].

Aesthetic evaluation of interfaces poses problem due to its subjective nature: an aesthetically pleasing interface may not look so to a different person. Computational aesthetic modeling attempts to overcome this problem by proposing objective measure of aesthetics [10, 16, 24].

One of the early works in this direction was by Ngo et al. [16]. In the approach, a numerical value is computed from the specification (in terms of elements, their positions, shapes and sizes) of an interface. The value signifies aesthetic of the layout. Aesthetics of two interfaces may be compared on the basis of the computed value. The model assumed a very simplified representation of the interface (i.e. each on-screen element is a rectangle). Aesthetic is determined by the geometric arrangement of the rectangles only. The content of the rectangles are not taken into consideration. Moreover, it considered only static images (i.e. the content does not change over time). Therefore, when we consider short videos embedded in a web page, it is not possible to apply the model, as we have to see "inside the box" (the content inside the rectangles) as well as consider the dynamic nature of the content.

In the context of short animation videos that are typically found embedded in web pages, some of these issues were addressed by Shyam and Bhattacharya [24]. In their work, a computational model was proposed to classify a short video into either of the classes *good, average* and *bad*, based on the aesthetic quality of the video. The model takes into consideration three factors that characterize a video, namely *symmetry*, *balance* and *color contrast*. We briefly discuss these factors in the following, as we have used them in our work.

The symmetry measure determines the extent to which the interface is symmetrical in vertical, horizontal and diagonal direction. In order to calculate symmetry (*Sym*) of an interface, Eq. 1 was proposed.

$$Sym = 1 - \frac{|S_h| + |S_v| + |S_r|}{3} \in [0,1] \quad \text{..... (1)}$$

In Eq. 1, S_h, S_v and S_r refers to the symmetry in horizontal, vertical and radial directions, respectively. Horizontal symmetry is calculated about a horizontal axis passing through the center of interface (Eq. 2a). Vertical symmetry is defined similarly, with respect to a vertical axis (Eq. 2b). Radial symmetry (Eq. 2c) refers to the symmetry about a diagonal passing through the center.

$$S_h = \frac{\begin{array}{l} |X'_{UL} - X'_{LL}| + |X'_{UR} - X'_{LR}| + \\ |Y'_{UL} - Y'_{LL}| + |Y'_{UR} - Y'_{LR}| + \\ |H'_{UL} - H'_{LL}| + |H'_{UR} - H'_{LR}| + \\ |B'_{UL} - B'_{LL}| + |B'_{UR} - B'_{LR}| + \\ |\theta'_{UL} - \theta'_{LL}| + |\theta'_{UR} - \theta'_{LR}| + \\ |R'_{UL} - R'_{LL}| + |R'_{UR} - R'_{LR}| \end{array}}{12} \quad \text{..... (2a)}$$

$$S_v = \frac{\begin{array}{l} |X'_{UL} - X'_{UR}| + |X'_{LL} - X'_{LR}| + \\ |Y'_{UL} - Y'_{UR}| + |Y'_{LL} - Y'_{LR}| + \\ |H'_{UL} - H'_{UR}| + |H'_{LL} - H'_{LR}| + \\ |B'_{UL} - B'_{UR}| + |B'_{LL} - B'_{LR}| + \\ |\theta'_{UL} - \theta'_{UR}| + |\theta'_{LL} - \theta'_{LR}| + \\ |R'_{UL} - R'_{UR}| + |R'_{LL} - R'_{LR}| \end{array}}{12} \quad \text{..... (2b)}$$

$$S_r = \frac{\begin{array}{l} |X'_{UL} - X'_{LR}| + |X'_{UR} - X'_{LL}| + \\ |Y'_{UL} - Y'_{LR}| + |Y'_{UR} - Y'_{LL}| + \\ |H'_{UL} - H'_{LR}| + |H'_{UR} - H'_{LL}| + \\ |B'_{UL} - B'_{LR}| + |B'_{UR} - B'_{LL}| + \\ |\theta'_{UL} - \theta'_{LR}| + |\theta'_{UR} - \theta'_{LL}| + \\ |R'_{UL} - R'_{LR}| + |R'_{UR} - R'_{LL}| \end{array}}{12} \quad \text{..... (2c)}$$

X'_j, Y'_j, H'_j, B'_j, θ'_j and R'_j (j = UR/UL/LR/LL) are the normalized values of the corresponding expressions shown in Eq. 3. UR, UL, LR and LL denote upper-right, upper-left, lower-right and lower-left, respectively. (x_{ij}, y_{ij}) and (x_c, y_c) in Eq. 3 refer to the center of each object *i* in quadrant *j* and the center of the interface. b_{ij} and h_{ij} are the width and height of the object. n_j is the total number of objects in the quadrant.

$$X_j = \sum_{i=1}^{n_j} | x_{ij} - x_c | \quad \text{(3a)}$$

$$Y_j = \sum_{i=1}^{n_j} | y_{ij} - y_c | \quad \text{(3b)}$$

$$H_j = \sum_{i=1}^{n_j} | h_{ij} | \quad \text{(3c)}$$

$$B_j = \sum_{i=1}^{n_j} | b_{ij} | \quad \text{(3d)}$$

$$\theta_j = \sum_{i=1}^{n_j} \frac{| y_{ij} - y_c |}{| x_{ij} - x_c |} \quad \text{(3e)}$$

$$R_j = \sum_{i=1}^{n_j} \sqrt{(x_{ij} - x_c)^2 + (y_{ij} - y_c)^2} \quad \text{(3f)}$$

The balance measure computes the difference between total optical weighting of components on each side of the horizontal and vertical axis. The optical weighting refers to the perception that some objects appear heavier than others. The expression for balance (*Bal*) is shown in Eq. 4, where B_h and B_v are the balance measured in the horizontal and vertical directions, respectively. Equation 5 shows the expressions to calculate the two components.

$$Bal = 1 - \frac{| B_h | + | B_v |}{2} \in [0,1] \quad \text{(4)}$$

$$B_h = \frac{w_T - w_B}{\max(| w_T |, | w_B |)} \quad \text{(5a)}$$

$$B_v = \frac{w_H - w_R}{\max(| w_H |, | w_R |)} \quad \text{(5b)}$$

In the above equations, $w_j = \sum_{i=1}^{n_j} a_{ij} d_{ij}$, j=L/R/T/B, L, R, T, B stand for left, right, top and bottom, respectively, a_{ij} is the area of object *i* on side *j*, d_{ij} is the

distance between the central lines of the object and the interface and n_j is the total number of objects on the side.

The above formulations were for static images. The idea is extended in [24] for video that is a sequence of frames. The symmetry and balance for each frame are calculated separately and then weighted averages of these individual values are calculated to get the respective symmetry and balance for the whole video, as shown in Eq. 6, where sym_i and bal_i are the symmetry and balance values of the i^{th} frame respectively, f is the total number of frames, sd_{ij} is the symmetry difference between consecutive frames and bd_{ij} is the balance difference between consecutive frames.

$$Sym = \frac{sym_1 + \sum_{i=2}^{f} sym_i \times \frac{1}{|\, sd_{i,i-1} \,|}}{1 + \sum_{i=2}^{f} \frac{1}{|\, sd_{i,i-1} \,|}} \in [0,1]$$

$$Bal = \frac{bal_1 + \sum_{i=2}^{f} bal_i \times \frac{1}{|\, bd_{i,i-1} \,|}}{1 + \sum_{i=2}^{f} \frac{1}{|\, bd_{i,i-1} \,|}} \in [0,1] \quad \text{…….. (6)}$$

Since the objects may change their position in a video, the above are calculated in terms of either *fixed objects* (i.e., those objects that don't change their position throughout the entire video) or the center of the frame (if there are no fixed objects).

The color contrast is the difference in visual properties that makes an object (or its representation in an image) distinguishable from other objects and the background. A three-stage approach was reported in [24] to calculate color contrast of a video. In the first stage, the video is divided into frames and then each frame is converted to gray image. Next, each gray image is converted to standard color enhanced image by histogram equalization. Finally, the original gray image is compared with the corresponding enhanced image in the third stage, to determine the color contrast of the video. Eq. 7 shows the computation of the color contrast (CC) value.

$$CC = \left| \frac{\sum_{i=1}^{p} std_i - org_i}{255 \times p} \right| \quad \text{……. (7)}$$

Based on these three factors, an expression shown in Eq. 8 was proposed in [24] to compute an aesthetic value (AS) for a video. On the basis of the computed value, videos are categorized as *good* (AS≥0.75), *average* (0.5≤AS<0.75) or *bad* (AS<0.5).

$$AS = \frac{Sym + Bal}{2} - CC \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Although high classification accuracy (about 87%) was reported by Shyam and Bhattacharya [24], the model was developed based on several assumptions. These include, (a) the objects in the videos are of regular shapes, (b) they do not change size across frames and (c) the objects follow linear motion paths. In this work, we propose a machine-learning based approach to overcome these limitations.

3 Proposed Model

We propose a classifier that is *trained* with a set of training videos (i.e., short videos that are already classified as *good, average* or *bad*). As it was difficult to create a large training set, we used the co-training algorithm [6], which can work on small training set. Co-training is a semi-supervised learning technique that requires two views (represented by two feature sets) of the data. Ideally, the two views are conditionally independent (i.e., the two feature sets are conditionally independent given the class) and each view is sufficient (i.e., the class of an instance can be accurately predicted from each view alone).

Co-training first learns a separate classifier for each view using a small set of labeled (training) examples. The most confident prediction of each classifier for an unlabeled data is then used to iteratively construct additional labeled training data. We used the *Naive Bayes* classifier [14] to classify data in the co-training method.

3.1 Identification of Feature Sets

The first step was the development of feature sets. A feature set denotes a set of features that characterize a short video. For the proposed model, we identified two feature sets, denoted by FS_1 and FS_2.

The feature set FS_1 contains three features, namely *symmetry*, *balance* and *color contrast*. These are the factors described in [24] (discussed in the related works section), each of whose value lies within the range [0, 1].

The feature set FS_2 was determined from a survey of 18 web pages, sampled randomly from the Internet, containing short videos. In the survey, we looked for the shapes (regular/irregular) of the objects in the video, motion pattern (linear/non-linear) of the objects, presence of *fixed* objects and change in object size across frames of a video. The observations are summarized in Table 1. From the table, we can conclude that the characteristics *object shape*, *change in size*, *presence of fixed objects* and *motion path* may have an influence on the perceived beauty (aesthetics) of a video. Along with those, it is also important to take into account the *total number of objects* in a video, since too many or too few objects may not be pleasing to the eye.

On the basis of the analysis of the survey results, we propose five features that form FS_2:

1. Total number of objects (N).

2. Fixed objects measure (represented as n_f/N, where n_f is the number of fixed objects).
3. Measure of size change across frames (represented as n_s/N, where n_s is the number of objects changing size).
4. Measure of movement path (represented as n_l/N, where n_l is the number of objects with linear movement).
5. Object shape measure.

Table 1. Summary of the observations made with 18 web pages with embedded short videos sampled from the Internet.

Characteristics	Observation
Object shape	Videos containing irregular shaped objects: 17 (94.4 %) Videos containing regular shaped objects: 1 (5.6%) Videos containing both regular and irregular shaped objects: 0 (0%)
Object size changes across frames	Videos where objects change their size across frame: 5 (27.8%) Videos where objects do not change their size: 13 (72.2%)
Fixed objects	Number of videos containing at least one fixed object: 7 (37.8%) Number of videos with no fixed objects: 11 (62.2%)
Motion paths	Videos containing objects with linear motion only: 6 (33.3%) Videos having objects with non-linear motion paths: 12 (66.7%)

In order to compute the last feature value (object shape), we used the formulation of Birkhoff [4], which works for object with polygonal shape[2]. According to the formulation, aesthetic quality (M) of any object can be computed in terms of *order* (O) and *complexity* (C) as in Eq. 9.

$$M=O/C \quad (9)$$

The Complexity C of an object is defined as the number of indefinitely extended straight lines which contain all the sides of the object (i.e., the number of distinct straight lines containing at least one side of the object). The Order O is a composition of five elements, as shown in Eq. 10.

$$O = V+E+R+HV-F \quad (10)$$

The individual terms on the right hand side of Eq. 10 are briefly described below (see [4] for more details).

- V stands for *vertical symmetry*. V=1 if the object posses symmetry about the vertical axis and V=0 otherwise.

[2] We can use this for any object shape in principle since any shape can be approximated with polygonal meshes. Thus, the solution is general, not specific to polygonal objects only.

- E stands for *equilibrium*. E=1 if V=1 or if the centre of the object is situated directly above a point P on a horizontal line segment AB supporting the object from below such that |AP| and |BP| > 1/6 of the total horizontal breadth of the object. If the center is above P but the above condition does not hold, E=0. For all other cases, E= -1.
- R stands for *rotational symmetry*. Let $360^o/q$ be the least degree of rotation which rotates the object into itself. Then, R=min{q/2,q/3} if V=1 for the object or its enclosing polygon, R=1 in any other case when q is even (i.e., in case of central symmetry) and R=0 otherwise.
- HV stands for relation of the object to a *horizontal-vertical network*. It can take the values of 0, 1 or 2 depending on the shape of the object.
- F stands for *unsatisfactory form*. F=0 if (a) the minimum distance from any vertex to any other vertex or side, or between parallel sides, is not less than $1/10^{th}$ the maximum distance between points of the polygonal object or (b) the angle between two non-parallel side is not less than 20^o or (c) there are at most two types of directions or (d) V and R are not both 0 or (e) there is at most one type of niche or (f) there is no unsupported re-entrant type. F=1 if the above conditions are fulfilled with only one exception. F=2 otherwise.

Let a video has f number of frames and M_{ij} is the Birkoff measure of the i^{th} object (i= 1,2...N) in the j^{th} frame (j= 1,2...f). Then, the object shape measure for the j^{th} frame (F_j) is computed as,

$$F_j = \frac{\sum_{i=1}^{N} M_{ij}}{N} \quad \text{..... (11)}$$

The above equation is for one frame. We calculate for each frame and take the average of all the frames. Hence, the object shape measure for the video is given as,

$$\text{Object shape measure} = \frac{\sum_{j=1}^{f} F_j}{f} \quad \text{..... (12)}$$

3.2 Creation of the Initial Training Set

The next step in the model development was the creation of a set of short videos that are already classified (i.e., labeled data). These labeled videos served as the initial training set. In order to create this training set, we conducted an empirical study in which we asked participants to rate a set of 18 artificially created short videos. From the participants' ratings, we labeled those 18 videos as *good*, *average* or *bad*. The details of the empirical study are discussed next.

Experimental Setup and Participants. We designed 18 videos using Adobe Flash Professional CS5™. The videos were divided into 3 sets of 6 videos each, containing regular shaped objects, irregular shaped objects and combination of both. We considered rectangular and circular shapes as regular. All other shapes were treated as irregular. Each video was displayed on a window of 320×233 resolution, had 40 frames with 2 sec duration (frame rate = 20) and were 2D, that is, the motion of all the objects were on a plane. The number of objects remained fixed in a video, that is, none of the objects were added or removed between the frames.

The total number of objects varied between 4 and 6 in each video. Two of the videos in each set contained fixed objects (1 and 3, respectively). One video in each set had 2 objects changing size across frames. The number of objects in linear motion varied between 0 and 4 in each set.

These 18 videos were shown to 17 participants on 17'' widescreen color displays attached to PCs having Intel® Core2™ Duo processor with 2.00 GHz speed, running Windows XP Professional with SP3. The participants included both male and female. All were either undergraduate or postgraduate students with average age of 21. All of them had normal or corrected-to-normal vision and were regular computer users. None were familiar with screen design concepts.

Table 2. Sequence of the videos in the playlist. Pi - playlist number.

P_1	1	2	18	3	17	4	16	5	15	6	14	7	13	8	12	9	11	10
P_2	2	3	1	4	18	5	17	6	16	7	15	8	14	9	13	10	12	11
P_3	3	4	2	5	1	6	18	7	17	8	16	9	15	10	14	11	13	12
P_4	4	5	3	6	2	7	1	8	18	9	17	10	16	11	15	12	14	13
P_5	5	6	4	7	3	8	2	9	1	10	18	11	17	12	16	13	15	14
P_6	6	7	5	8	4	9	3	10	2	11	1	12	18	13	17	14	16	15
P_7	7	8	6	9	5	10	4	11	3	12	2	13	1	14	18	15	17	16
P_8	8	9	7	10	6	11	5	12	4	13	3	14	2	15	1	16	18	17
P_9	9	10	8	11	7	12	6	13	5	14	4	15	3	16	2	17	1	18
P_{10}	10	11	9	12	8	13	7	14	6	15	5	16	4	17	3	18	2	1
P_{11}	11	12	10	13	9	14	8	15	7	16	6	17	5	18	4	1	3	2
P_{12}	12	13	11	14	10	15	9	16	8	17	7	18	6	1	5	2	4	3
P_{13}	13	14	12	15	11	16	10	17	9	18	8	1	7	2	6	3	5	4
P_{14}	14	15	13	16	12	17	11	18	10	1	9	2	8	3	7	4	6	5
P_{15}	15	16	14	17	13	18	12	1	11	2	10	3	9	4	8	5	7	6
P_{16}	16	17	15	18	14	1	13	2	12	3	11	4	10	5	9	6	8	7
P_{17}	17	18	16	1	15	2	14	3	13	4	12	5	11	6	10	7	9	8

Data Collection Procedure. We created 17 playlists for 17 participants using *balanced Latin squares* [5] (see Table 2). In Table 2, each row represents a playlist (P_i) shown to the i^{th} participant. The numbers in each cell represent one of the 18 videos. Video numbers 1-6 correspond to the set containing regular objects, 7-12 correspond to the set containing irregular objects and 13-18 correspond to the set containing both types of objects.

The videos were shown to the participants in the sequence shown in Table 2 and they were asked to rate the videos on a scale of 1 (least attractive) - 7 (most attractive) as per their perception of the attractiveness of the videos. Figure 1 shows the screenshot of the interface used by the participants to rate the videos. A play button allowed the participant to play the next video in the list, once s/he was finished with the current video. A replay button was also provided to enable the participant replay the current video. In the figure, it can be seen that the entire background screen was covered by the interface while the participant was rating the videos. This was done to avoid distraction.

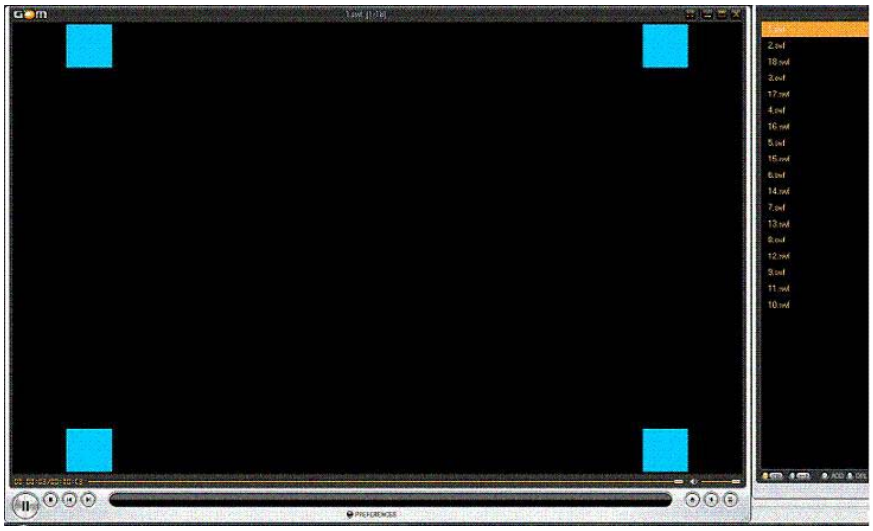

Fig 1. Screenshot of the rating interface.

The ratings by the participants are shown in Table 3. We mapped the participants' rating to one of the three classes *good*, *average* and *bad*. We considered a rating of 1, 2 and 3 as *bad*, 4 and 5 as *average* and 6 and 7 as *good*. After the mapping, we took the statistical *mode* of the classes for each video, which was the final label (class) of the video. In case of a tie (i.e., more than one class occur in equal number), we take the average of the original ratings. The average value was used to assign class (between 1-3 as *bad*, 4-5 as *average* and 6-7 as *good*). The results are summarized in Table 4. From Table 4, it can be seen that three videos were labeled as *bad*, five were labeled as *good* and the remaining ten were labeled as *average*.

Table 3. Rankings of the videos by participants (in the scale of 1-7). The numbers in the top row denotes the videos and the leftmost column shows the participants.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
$\mathbf{P_1}$	2	3	5	6	4	7	5	2	5	4	3	7	5	5	7	6	2	5
$\mathbf{P_2}$	4	4	5	6	5	6	5	5	5	6	5	7	5	5	5	6	5	7
$\mathbf{P_3}$	2	3	3	5	4	5	2	3	4	4	3	5	4	4	4	5	4	6
$\mathbf{P_4}$	3	4	5	7	3	7	4	3	5	3	4	6	5	2	5	2	3	4
$\mathbf{P_5}$	5	5	5	6	2	7	4	4	4	5	3	7	2	4	2	6	3	7
$\mathbf{P_6}$	5	4	4	5	4	5	3	4	5	5	4	6	4	5	5	6	4	6
$\mathbf{P_7}$	7	6	4	7	5	6	4	2	3	3	4	4	2	2	3	2	2	3
$\mathbf{P_8}$	2	4	4	2	4	5	4	4	4	6	2	7	5	5	4	6	1	6
$\mathbf{P_9}$	3	3	3	5	4	5	4	4	7	5	4	7	4	4	5	5	4	7
$\mathbf{P_{10}}$	6	4	4	7	5	6	4	4	3	6	4	5	4	4	5	6	5	6
$\mathbf{P_{11}}$	2	4	3	4	3	4	4	4	6	5	5	6	5	6	6	7	7	7
$\mathbf{P_{12}}$	4	3	4	5	4	6	4	3	5	5	4	5	3	4	3	4	3	6
$\mathbf{P_{13}}$	4	3	5	2	2	7	6	4	3	2	1	3	2	3	4	2	1	6
$\mathbf{P_{14}}$	1	3	4	5	7	6	5	4	5	3	6	5	7	5	6	7	6	6
$\mathbf{P_{15}}$	4	1	3	2	6	7	2	4	5	7	6	2	1	4	1	2	3	5
$\mathbf{P_{16}}$	1	1	2	6	5	7	4	5	3	7	7	3	2	2	1	2	2	5
$\mathbf{P_{17}}$	2	4	5	3	6	6	4	3	1	3	4	6	7	5	4	7	6	5

3.3 Unlabeled Dataset Creation

We designed another set of 24 short videos, which served as the unlabeled dataset (i.e., these were not classified from empirical data), using the same development platform as that of the labeled videos. The resolution, frame rate and duration of the videos were also the same along with the nature of the motion paths of the objects (2D).

The purpose of these unlabeled videos was to increase the training set size so as to cover a wide range of values for all the features. The videos were divided into 3 sets of 8 videos each. One set contained videos with regular objects only, one set was having videos with only irregular objects and the third set was having videos containing both regular and irregular objects.

The total number of objects in the videos varied between 2 to 7. The videos contained between 0 (5 videos) and 3 fixed objects. About 50% of the videos (13) contained objects (between 1 and 3) that changed size across frames. Two of the videos did not have any objects following linear motion path. In the remaining 22 videos, objects with linear motion path varied between 1 and 5.

Table 4. Labeling of videos from participants' rating (Table 3). The numbers inside parenthesis in the middle column show the number of participants who rated the video to belong to the corresponding class. A rating of 1, 2 or 3 was mapped to *bad*, 4 or 5 was mapped to *average* and 6 or 7 was mapped to *good* class. The final label is obtained as the statistical mode of the labels given by the participants.

Video	Participant rating	Final Label
1	Bad (9 participant), Average (6 participant), Good (2 participant)	Bad
2	Bad (8 participant), Average (8 participant), Good (1 participant)	Bad
3	Bad (5 participant), Average (13 participant), Good (0 participant)	Average
4	Bad (4 participant), Average (6 participant), Good (7 participant)	Good
5	Bad (4 participant), Average (10 participant), Good (3 participant)	Average
6	Bad (0 participant), Average (5 participant), Good (12 participant)	Good
7	Bad (3 participant), Average (13 participant), Good (1 participant)	Average
8	Bad (6 participant), Average (11 participant), Good (0 participant)	Average
9	Bad (5 participant), Average (10 participant), Good (2 participant)	Average
10	Bad (5 participant), Average (7 participant), Good (5 participant)	Average
11	Bad (5 participant), Average (9 participant), Good (3 participant)	Average
12	Bad (3 participant), Average (5 participant), Good (9 participant)	Good
13	Bad (6 participant), Average (9 participant), Good (2 participant)	Average
14	Bad (4 participant), Average (12 participant), Good (1 participant)	Average
15	Bad (5 participant), Average (9 participant), Good (3 participant)	Average
16	Bad (5 participant), Average (3 participant), Good (9 participant)	Good
17	Bad (9 participant), Average (5 participant), Good (3 participant)	Bad
18	Bad (1 participant), Average (5 participant), Good (11 participant)	Good

3.4 Implementation of the Training Method

The implementation of the Co-training method was done in MATLAB™. In order to calculate the feature values in the feature sets FS_1 and FS_2, we first divided a video into frames or sequence of images. Then, we tracked objects in each frame and found out the coordinates of the center of every tracked object. For tracking the objects in each frame, we first converted the frame to a binary image. Then, we applied the *bwmorph* function, which shrinks the objects to points. The final frame contains only points representing the number of objects in the frame. The steps are illustrated in Fig. 2(a)-(c).

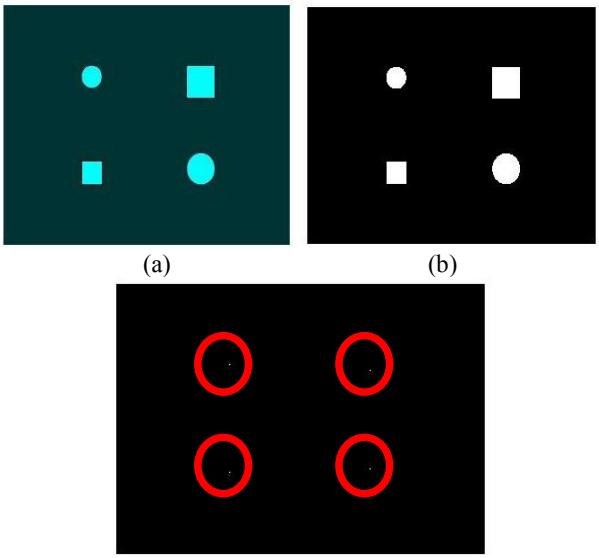

Fig 2. Illustration of the feature value computation steps. From the frame in (a), a binary image is created (b). It is then converted to points representing objects (c) (shown inside circles).

Using the co-training algorithm, we classified the unlabeled videos. Among those videos, 7 were classified as *bad*, 11 as *average* and the remaining 6 as *good*.

4 Model Validation

We conducted an empirical study to check the accuracy of the model (classifier). In the study, we used the model to classify 16 short videos. The videos were then rated by 23 participants. From the rating, we determined the classes of the videos. The model classifications were then matched with the empirical classification to determine model accuracy. The details of the validation study are described next.

4.1 Experimental Setup

All the 16 videos were 2D (i.e., objects moved in 2D), designed using Adobe Flash Professional CS5™ as before. Other characteristics, namely the frame rate, display resolution, total number of frames and duration were also the same as to that of the videos designed for training the model. None of the objects in a video was added or removed during the running of the video. The feature values were varied at random in the videos.

4.2 Participants and Procedure

Among the 23 participants, 10 took part in the previous study and 13 were new. All the participants were undergraduate or postgraduate students with regular computer exposure. Average age of the participants was 21.23 yrs. Among them, 15 were male and the rest were females. All of them had normal or corrected to normal vision. None of them had any experience with screen-design concepts before.

In order to collect data, the procedure we followed was similar to the one we used for labeling of the training videos. We created playlists of the videos for each participant following the Latin square method. The participants were asked to rate the videos using the same interface and rating scale. The ratings were then mapped to one of the classes, leading to the statistical mode based final classification of the videos.

4.3 Results

According to the participants' ratings, 4 videos were classified as *bad*, 4 as *good* and the remaining 8 as *average*. The classification we obtained using the model matched 12 of these empirical classes, resulting in 75% accuracy. The results of the study are summarized in Table 5.

Table 5. The comparison of the model prediction to that of the classification from empirical data.

Video Number	Empirical Classification	Model Prediction
1	Bad	Bad
2	Bad	Bad
3	Good	Good
4	Average	Average
5	Good	Average
6	Bad	Average
7	Average	Average
8	Good	Good
9	Bad	Bad
10	Average	Average
11	Average	Good
12	Good	Good
13	Average	Average
14	Average	Bad
15	Average	Average
16	Average	Average

5 Discussion

In this work, we tried to address the limitations in the work reported by Shyam and Bhattacharya [24], by proposing a more generalizable classifier, which is trained using the co-training method. The results of the validation study show that the proposed classifier is able to classify short videos according to their aesthetic appeal, with a reasonably high accuracy rate of 75%.

The classification helps a designer decide if a video needs to be improved to increase its aesthetic appeal. For videos belonging to the *good* category, improvements may not be necessary. For *average* category videos, improvements may help while for videos classified as *bad*, it is definitely required. As is obvious, this has significant implication from the point of view of usability of web pages, when we consider web pages with embedded videos. We believe the work can be extended for the development of a more generalized aesthetic model for web pages.

An important characteristic of the videos used in the study was that the number of objects remained fixed (i.e., no addition/deletion of objects was considered). Admittedly, the constraint may not characterize some real-world embedded videos. Therefore, it may be necessary to carry out further work to determine the validity of the proposed model for videos that do not have fixed number of objects.

Although the model accuracy was reasonably satisfactory, we feel that further improvements are possible. The feature sets were developed on the basis of a survey of 18 videos sampled from the Internet. A larger sample size may reveal other characteristics, thereby enriching the feature set. Moreover, the initial training set was created with data of 17 participants for 18 videos. There are scopes to improve the initial training set by increasing the number of videos and participants and also by introducing more variations, in terms of age, gender, educational background and so on, to the participants' profile. Finally, the accuracy figure also needs to be corroborated further by considering more videos and larger number of participants with more variations in their profile.

6 Conclusions

In this paper, we reported a computational model to classify short videos based on their aesthetic quality. The model is a Naïve Bayes classifier, developed using the co-training method. The model was developed and validated using empirical data. Experiments show that the model can classify videos with 75% accuracy.

In future, we plan to work on two directions: refinement of the model and using the model to propose an overall computational model for aesthetic evaluation of a web page. We plan to refine the model by carrying out the following tasks.

- Refinement of the feature set by surveying larger number of real-world embedded videos.
- Increase the initial training set by increasing the number of videos and a larger set of participants with more varied profile to label those videos.

- Perform more extensive validation experiments with more videos and larger number of participants with more variations in their profile.

Acknowledgements. We are thankful to all the participants who volunteered for the empirical studies.

References

1. Arnheim. R. Art and Visual Perception. University of California Press, Berkeley, CA, New York, 1954.
2. Aspillaga, M. Screen design: Location of information and its effects on learning. Journal of Computer-Based Instruction (1991), 89-92.
3. Bartelsen, O. W., Petersen, M. G., Pold, S. (eds). Aesthetic Approaches to Human-Computer Interaction, NordiCHI 2004 Workshop (2004), Tampere, Finland.
4. Birkhoff, G. D. Aesthetic Measure. Chapter 1, Harvard University Press, Cambridge, MA., 1933.
5. Blandford, A., Cox, A. L. and Cairns, P. Controlled experiments. In P. Cairns and A. L. Cox, editors, Research Methods for Human-Computer Interaction, pages 1–16. Cambridge University Press, 2011.
6. Blum, A., Mitchell, T. Combining Labeled and Unlabeled Data with Co-training, In Proc. 11th Annual Conference on Computational Learning Theory, pp 92-100, 1998.
7. Dondis, D. A. A Primer of Visual Literacy. The MIT Press, Cambridge, MA, 1973.
8. Galitz, W. O. The essential guide to user interface design: an introduction to GUI design principles and techniques. John Wiley Sons Inc, New York, 1997.
9. Heines, J. Screen Design Strategies for Computer-assisted Instruction. Digital Press, Bedford, MA, 1984.
10. Lai, C. Y., Chen, P. H., Shih, S. W., Liu, Y., Hong, J. S. Computational models and experimental investigations of effects of balance and symmetry on the aesthetics of text-overlaid images. International Journal of Human-Computer Studies, 68 (2010), 41–56.
11. Lavie, T., Tractinsky, N. Assessing dimensions of perceived visual aesthetics of web sites. International Journal of Human-Computer Studies, 60, 3 (2004), 269–298.
12. Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., Noonan, L. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. ACM Transactions on Computer-Human Interaction, 18, 1 (2011).
13. Maquet, J. The Aesthetic Experience. Yale University Press, New Haven, CT, 1986.
14. Mitchell, T. Machine Learning. Chapter 6, McGraw-Hill, 1997.
15. Moshagen, M., Thielsch, M. T. Facets of visual aesthetics. International Journal of Human-Computer Studies, 68 (2010), 689–709.
16. Ngo, D. C. L., Teo, L. S. and Byrne, J. G. Modeling interface aesthetics. Information Sciences, 152 (2003), 25–46.
17. Norman, D. A. Introduction to this special section on beauty, goodness, and usability. Human-Computer Interaction, 19 (2004), 311–318.
18. Pandir, M., Knight, J. Homepage aesthetics: The search for preference factors and the challenges of subjectivity. Interacting with Computers, 18, 6(2006), 1351–1370.
19. Park, S., Choi, D., Kim, J. Critical factors for the aesthetic fidelity of web pages: empirical studies with professional web designers and users. Interacting with Computers, 16, 2 (2004), 127–145.

20. Petersen, M. G., Hallinas, L. and Jacob, R. J. K. Introduction to special issue on the aesthetics of interaction. ACM Transactions on Human-Computer Interaction, 15, 4 (2008).
21. Reilly, S., Roach, J. Improved visual design for graphics display. IEEE Computer Graphics and Applications, 4, 2 (1984), 42–51.
22. Schaik, P. V., Ling, J. Five psychometric scales for online measurement of the quality of human-computer interaction in web sites. International Journal of Human-Computer Interaction, 18, 3 (2005), 309–322.
23. Schmidt, K., Liu, Y., Sridvasan, S. Webpage aesthetics, performance and usability: Design variables and their effects. Ergonomics, 52, 6 (2009), 631–643.
24. Shyam, D., Bhattacharya, S. A Model to Evaluate Aesthetics of Short Videos. In Proc. 10th Asia Pacific Conference on Computer Human Interaction (APCHI 2012), Matsue, Japan, 2012, pp. 315-324.
25. Szabo, M., Kanuka, H. Effects of violating screen design principles of balance, unity and focus on recall learning, study time, and completion rates. In Proc. ED-Media/ED-Telecom 98 Conference, Charlottesville, VA, 1998. Association for the Advancement of Computing in Education.
26. Toh, S. C. Cognitive and Motivational Effects of Two Multimedia Simulation Presentation Modes on Science Learning. PhD thesis, University of Science Malaysia, Malaysia, 1998.
27. Tractinsky, N. Aesthetics and apparent usability: Empirically assessing cultural and methodological issues. In CHI ’97 Conference Proceedings, New York, 1997. ACM.
28. Tractinsky, N. Does aesthetics matter in human computer interaction? Mensch and Computer (2005), 29–42.
29. Tractinsky, N., Hassenzhal, M. Arguing for aesthetics in human-computer interaction. i-com Z. Interakt. Koop. Medien (2005), 66–68.
30. Tractinsky, N., Shoval-Katz, A. and Ikar, D. What is beautiful is usable. Interacting with Computers, 13, 2 (2000), 127–145.
31. Tullis, T. S. An evaluation of alphanumeric, graphic, and colour information displays. Human Factors, 23 (1981), 541–550.
32. Tullis, T. S. Predicting the Usability of Alphanumeric Displays. PhD thesis, Rice University, Kansas, 1984.
33. Tullis, T. S. Screen design. In M. Helander, editor, Handbook of Human-Computer Interaction, pages 377–411. Elsevier Science Publishers, Amsterdam, the Netherlands, 1988.