Authorship Detection and Encoding for eBay Images

Liping Zhou, Wei-Bang Chen, Chengcui Zhang
Department of Computer and Information Sciences,
The University of Alabama at Birmingham, AL 35294 USA

ABSTRACT

This paper describes a framework to detect authorship of eBay images which contains three modules - editing style summarization, classification and multi-account linking detection. For editing style summarization, three approaches, namely the edge-based approach, the color-based approach, and the color probability approach, are proposed to encode the common patterns inside a group of images with similar editing styles into common edge or color models. Prior to the summarization step, in order to group images with similar editing styles together, an edge-based clustering algorithm is developed for this purpose. Corresponding to the three summarization approaches, three classification methods are developed accordingly to predict the authorships of an unlabeled test image. For multi-account linking detection, in order to detect the hidden same owner behind multiple eBay seller accounts, two methods to measure the similarity between seller accounts based on similar models are presented. Experiments show promising results of our proposed framework.

Keywords: multimedia applications; authorship detection; image editing style; image clustering; classification

INTRODUCTION

Authorship detection has a range of applications in a large number of fields such as forensic evidence, plagiarism detection, email filtering, and web information management (Chen, S.C., 2010; Love, 2002; Rafailidis, Nanopoulos & Manolopoulos, 2010). In recent years, an application with growing interests is web information management. The World Wide Web provides a powerful publication platform, where a large number of images are created for different purposes. For instance, online shopping websites such as eBay enable sellers and buyers to transact on the platform for a broad variety of goods and services worldwide. The goods/products are mostly visualized by having their pictures displayed on the website so that customers can see the visual details of products which often play an important role in customers' purchase decision-making. EBay sellers, especially those power sellers, often add their own touch to product image design in order to attract buyers' attention. For example, some sellers add a frame and/or some promotion texts to their listing images (see Figure 1), and some sellers embed the name of their store as watermarks or logos in their images. We assume that many eBay sellers have developed distinctive editing styles over time to embed to their product images (logos, background, and so on.). Such editing styles are highly repetitive within one seller's images, but mostly distinctive among different sellers. We detect and encode such editing styles for each seller using visual features extracted from product images, and in turn use the encoded editing styles to automatically predict the ownership for unlabeled images. Through a collaborative effort between the authors' institution and eBay, the output of this study will be used as added clues in eBay's seller profiling system to detect and prevent account taken-over and other related fraudulent behaviors.

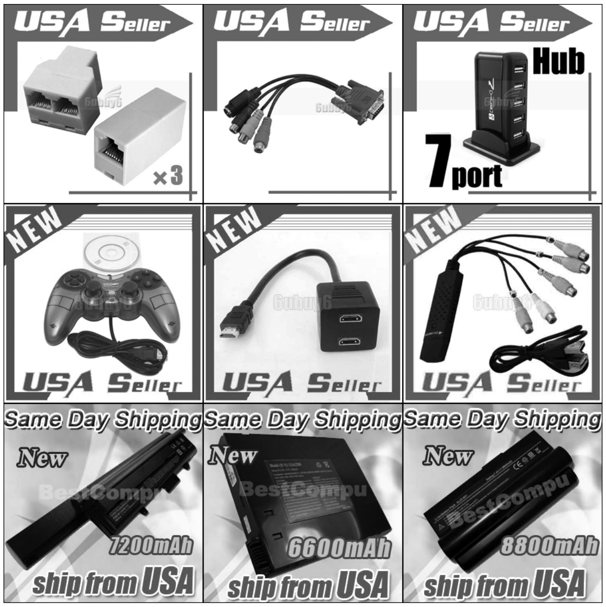

Figure 1. Examples of product images from online shopping websites such as eBay

There are several challenges in this study. First, we found that sellers may use more than one image editing style in composing their product images posted on eBay. For example, in Figure 1, the images in the first two rows all belong to the same seller "6ubuy6" but apparently they have visually different editing styles. Therefore, a simple image averaging technique applied to the images within the same seller will not work in the presence of multiple editing styles. Another challenge is that the same product image can be used by multiple sellers who re-edit the image according to their own style, and thus clustering based on global visual features will generate a large amount of false positives due to the common features extracted from the product itself. In this paper, we present a new edge-based clustering method to divide all images within one seller into image groups each of which corresponds to one editing style of that seller in terms of image edge maps (Abdel-Mottaleb, 2000). The clustering algorithm is based on the similarity of image edge maps. After that, summarization can be done to find the common pattern in each image group.

In order to summarize editing styles, we present three summarization methods: 1) edge-based summarization by applying Hough transform (Fernandes & Oliveira, 2008) on edge map images, which generates an edge template for each image group; 2) color-based summarization in HSV color space (Zhao, Bu & Chen, 2002) by using a mean image to represent the centroid of an image group, resulting in a HSV color model for each group; 3) color probability based summarization which generates a probability model to represent the spatial color distribution within an image group. The templates or models generated from the summarization step are then used to identify the authorship of unlabeled images using the corresponding classification methods which are edge-based classification with Hough transform, color-based classification, and color probability based classification. Based on our experiments, edge-based classification with Hough transform performs the best in term of accuracy. However, it is more time-consuming than the other two methods. Color-based classification by using overlay method in HSV color space is the most efficient among the three methods, but its accuracy is lower than that of the edge-based method. Since the three methods are based on different features, our guess is that they may be able to complement each other, which is actually evidenced by our experimental results. In fact, by combining the edge features and color features in our classification method, the performance has been improved, and the overall accuracy goes up to 94%.

In this framework, we label images with the predicted seller ID based on their image editing styles. However, some sellers may have multiple accounts (IDs), and they tend to use the same or similar editing styles in all of their accounts. Some of these sellers are indeed honest sellers who register for multiple accounts for the purpose of either self-promoting (e.g., giving positive feedbacks to each other to establish a high reputation) or selling different categories of products (rarely seen), but some others are real fraudsters who use multiple accounts to conduct self-promoting and fraudulent transactions. In either case, being able to know the linking between and among seller accounts would be very valuable information for eBay to fight online fraud. In this study, we assume that if the images from two seller accounts have a very similar editing style, then they probably belong to the same seller. By summarizing and encoding the image editing style(s) of each seller with image editing templates, we can discover the linking between sellers through the similarity analysis of their image editing templates. In particular, in this study, we propose two methods to indentify similar editing styles from across sellers, namely the covariance method and the clustering based method.

The arguments for our framework start with a brief discussion on authorship detection and encoding framework. Next, the implementation of proposed framework is followed by the discussion of the edge-based clustering algorithm, the three summarization methods including an edge-based summarization method and two color-based summarization methods and three classification methods. Thirdly, two multi-account linking detection approaches are described. Finally, the evaluation of system performance with experimental results is presented, followed by the summary and conclusion.

RELATED WORK

Authorship detection has been widely applied in electronic text, which identifies the author of an anonymous text, or the text whose authorship is in doubt. Due to the vast electronic text that have become available on the Internet recently, authorship detection techniques have become to play an increasingly significant role in areas such as document categorization, plagiarism analysis and near-duplicate detection (Zhao & Zobel, 2005). However, the dimension of multimedia analysis is largely missing in these efforts.

Robine, Hanna, Ferraro, and Allali (2007) reported a method to detect near-duplicate music documents and plagiarisms, where the similarity of two music documents is evaluated by the similarity of a pair of musical segments. In recent years, image authorship detection becomes an interesting application domain. Shun and Mitsu (2008) described a method of identifying authorship of Ukiyoe prints by using Rakkan images found in the prints, which is the seal, or signature included in the paintings. In their approach, the distance between dictionary templates and test data is calculated in order to identify artists and creation dates of Ukiyoe prints.

Watermarking is a commonly used approach for image authentication. However, existing watermark detection cannot be directly applied to eBay images for authorship detection due to the following reasons. First, although human can perceive visible watermarks on an image, the embedded pattern of a visible watermark should be difficult or even impossible to be removed unless intensive and expensive human labors are involved (Huang & Wu, 2004). In Huang and Wu (2004), a visible watermarking attacking scheme is proposed which manually selects watermarked areas and applies image recovery techniques to remove watermarks. However, this approach is impractical for analyzing a large-scale data set such as eBay images, especially given that many sellers do not have watermarks but exhaustive manual check cannot be avoided. There have been several other attempts to automatically extract watermarks by adopting Independent Component Analysis (ICA) (Pei & Zeng, 2006; Yu, Sattar & Ma, 2002). Performing ICA requires multiple observations of the source signals including observations for at least the original image and the mix of the original image and the watermark. However, the original image is not readily available for eBay images with watermarks. In Pei, et al (2006), the paper proposes to use manual intervention to generate multiple observations. However, again this is not practical for analyzing a large-scale data set such as eBay images. Further, watermark detection is not applicable to the images of those sellers who do not use watermarks in their editing templates.

In this research, an innovative framework based on image editing style analysis is proposed for online listing image authorship identification. Like the work presented in Hirose, Yoshimura, Hachimura, and Akama (2008), we look for “signatures” of an authorship rather than the global visual features in an image for the reason mentioned in previous section. Specifically, the framework is able to discover all editing styles created by a seller in an automatic manner by clustering the images owned by that seller on the basis of the edge map similarity. Ideally, each cluster produced corresponds to a distinctive editing style created by that seller. Therefore, all images in a cluster can be further encoded into a template (model) that represents an editing style either based on their common edge maps or common HSV color features. When predicting the authorship of an unlabeled image, the proposed framework measures the similarity between the unlabeled image and each template produced. A higher similarity value indicates more likely that the unlabeled image may have the same editing style as the template, and therefore, the image is more likely to belong to the seller who creates the template.

EDGE-BASED CLUSTERING

In this section, we present an edge-based clustering algorithm to divide images of each seller into different image groups based on their editing styles. Edge features are chosen because they are more robust than color features which often vary according to re-editing, re-compression, and change of fill-in colors. For example, in Figure 1, the three images in the second row belong to the same seller and have almost identical editing styles, but their frame colors are quite different.

In most images, the product usually occupies a large area; therefore, clustering based on global edge features will incorrectly cluster those images with common edges on the product area into the same group. To reduce the negative impact of edges on products, we only focus on the edges with strong intensities since image areas with added editing effects usually have strong edges, and also we try to ignore the edges in a fixed area around the center of the image, because products are almost always located in the center of images.

To calculate the similarity between two edge maps E_1 and E_2, we apply Generalized Hough transform (GHT) which is essentially a method originally used for object recognition. The basic process starts by assuming that one of the two edge maps has the predefined object. And the edge points from another edge map are mapped to the parameter space, which represents all the possible instances of edge features in the predefined object. Each matched edge point votes for the location of the predefined object, and the instance with the most votes defines the features present in the edge map, which defines the best matched points of two edge maps. Our proposed similarity measure further decides how likely an edge map has the predefined object. Based on the best matched points, we do the overlay of two edge maps to find the common area of two edge maps. The common pixels are defined as the pixels in the corresponding positions of two edge maps which have the same 1 or 0 value (edge or non-edge) within the common area. Finally, the ratio of the number of common pixels to the common area is used to measure the similarity of edge maps. The similarity of two edge maps is defined in Eq. 1 where C denotes their common area after image overlay, area(C) denotes the area of C, and $sum(pix)$ represents the number of common pixels in the two edge maps.

$$Similarity(E_1, E_2) = \frac{sum(pix)}{area(C)} \qquad (1)$$

A threshold is used in the clustering. If the similarity is greater than the threshold value, we consider the two images to be in the same group. The threshold value is experimentally set to 0.8 in this study.

SUMMARIZATION METHODS

After clustering, the images of each seller are clustered into image groups each of which represents one distinctive editing style of that seller's. In this section, based on edge and color features of images, three summarization methods are proposed to encode the editing styles with templates or models.

Edge-based Summarization

In this method, we summarize each image group with an edge template by finding the common edges of images. After the clustering step, the images of each seller are clustered into image groups based on the similarity of their edge maps; therefore, the images in each group must have similar edge features. These image groups are used as the training set to generate edge templates. In this process, edge detection is first performed to extract the edge map features for each image in an image group according to the clustering results, and then image overlay is conducted based on the best matched points which are generated in the clustering step. In the new edge template, each pixel value is the average of all pixel values (edge intensities) at that same location in the edge maps of all the training images in that group.

In order to reduce the noise, we choose to use binary values to describe the edge template, that is, the pixel value under 0.5 is set to 0, and the pixel value equal to or greater than 0.5 is set to 1. Further, if one image group has only one or two images, there is not enough training data for summarizing the common edge pattern in that image group. Therefore, we ignore the image groups containing less than 3 images.

The steps to generate an edge template are summarized as follows:

1. Convert color images in a group to grayscale images.
2. Use Sobel filter (Jahne, Scharr & Korkel, 1999) for edge detection and extract edge maps for the images.
3. Convert each edge map image into intensity image, and keep only those edges with strong intensities.
4. Compute the average value of each pixel, i.e., its average edge intensity, across all the edges maps.
5. Convert the edge intensity value of each pixel into a binary value to generate a new edge template.

Color-based Summarization

In the edge-based summarization method, we summarize and encode image editing styles of sellers into a set of edge templates by applying Hough transform on edge maps. Though image matching with Hough transform performs well, it is time-consuming. Therefore, based on the same edge-map clustering results, we also design two alternative color-based approaches to summarize common color patterns in HSV color space. HSV color space is chosen because it is more perceptually uniform.

Based on our clustering algorithm, we assume that images in each image group should have the same or similar editing style, so there is less distinction of color between images in the same image group but more distinction between images from different image groups. According to this assumption, we compute a mean image for each image group, and then a template can be represented by a sequence of pixel-level mean values.

An alternative color-based method is based on the assumption that all images associated with a template have perceptually similar color and layout because they have the same or similar editing style, and all color values associated with a pixel at a specific location in all training images may form a normal distribution. This pixel-level color distribution can be further represented as the mean and the standard deviation of all the color values at that pixel location. In this way, a color template can be described by a set of pixel-level color distributions as a probability model.

In both methods, all images are converted from RGB color space to HSV color space. In order to reduce the computational complexity, we have tested various image down-sampling rates in our experiments. The experimental results suggest that the performance has no significant difference between down-sampling to 200×200 pixels and 100×100 pixels. Thus, we down-sampled all images to 100×100 pixels in this study. The model generated from an image group is also a

100×100 image. In addition, to reduce the influence of product areas, we remove a central region with a fixed ratio of area from each image.

CLASSIFICATION METHODS

In order to predict the authorship of an unlabeled image, we implement three classification methods that correspond to the three summarization methods, respectively. The edge templates and color models generated in the summarization step are further used to identify the authorship of unlabeled images. More specifically, each test image is compared against each template/model and the best matched model/template will be used to predict the authorship for that image. In other words, a test image will be associated with the seller one of whose editing styles (models/templates) has the best match with the test image.

Edge-based Classification

In this method, we match the edge map extracted from a test image with each edge template generated from the summarization step. The edge map of the test image is also generated by using Sobel filter. The proposed algorithm then uses Hough transform to find the best matched point in the edge map of the test image when it is matched with an edge template. The calculation of similarity scores is similar to the method mentioned in the edge-based Summarization Method. Then the edge template which has the maximum similarity with the test image will be considered as its matched template. In our dataset, some images do not have any editing style, so they usually have a low similarity with each edge template. In order to avoid incorrectly associating such images with a template, we define a cutoff value. Therefore, if the maximum similarity score received by a test image is less than the cutoff value, that image cannot be matched with any template.

Color-based Classification

Using color-based summarization method, we summarize the color patterns in an image group to a mean color model in HSV color space. The similarity between a test image and a mean color model in HSV space is calculated as Eq. 3. For each pixel in the test image, we calculate the distance between its pixel value and the corresponding pixel value in a color model, and the sum of the distances of pixel values is used to represent the relative distance between the test image and the model. The distance between a pair of pixels can be computed in Eq. 2 where H_{test} is the "Hue" value of the pixel in the test image; H_{model} is the "Hue" value of the pixel in the model. S_{test} denotes the "Saturation" value of the pixel in the test image and V_{test} is the "Value" value of that pixel in the test image.

$$\mathrm{dist}(pix_{test}, pix_{model}) = (H_{test} - H_{model})^2 + (S_{test} - S_{model})^2 + (V_{test} - V_{model})^2 \qquad (2)$$

$$\mathrm{DIST}(I_{test}, I_{model}) = \mathrm{sum}(\mathrm{dist}(pix_{test}, pix_{model})) \qquad (3)$$

In this method, we also define a cutoff value to identify the images without any editing style. Therefore, if the relative distance between the test image and a model is greater than the cutoff value, the test image cannot be matched with that model.

Probability Model Classification

An alternative color-based approach for the image authorship classification is color probability based classification. For this purpose, we build a color probability model for each image group. Each model describes the pixel-level color distribution of all the images in that group. When performing the authorship identification, we measure the likelihood a test image belongs to each model.

In this method, given a pixel location, after collecting all the corresponding pixel values at that location from all images in the group, we can simply use z-test (Sprinthall, 2002), a statistical hypothesis test, to tell whether a color value comes from the same color distribution at a particular location, and thus, can predict whether a pixel's value belongs to the same pixel value distribution at the corresponding location in the template. This idea can be extended to test all pixels in a test image against a color probability model. Therefore, we define the overall likelihood between a test image and a color probability model as the mean probability of all pixels.

A cutoff value is used to again identify those images that do not have significant editing styles. More specifically, only those likelihood values exceeding our predefined cutoff values will be collected. A test image will be assigned to the seller who owns the model which yields the maximum likelihood for that image among all models.

EXPERIMENTS

The proposed framework is applied on two datasets in order to evaluate the performance of the proposed system. The first dataset (Dataset-I) consists of 919 images from 10 sellers, and the second dataset (Dataset-II) consists of 3980 images from 47 sellers. In Dataset-I, images procured from 7 sellers have significant editing styles while the rest of the images collected from the other 3 sellers do not have an obvious editing style. In Dataset-II, 7 sellers do not use any editing style in their images while the other 40 sellers create at least one editing style in their images.

In our experiments, a 10-fold cross-validation is performed on Dataset-I and a 3-fold cross-validation is performed on Dataset-II. The reason we did not perform 10-fold cross-validation on Dataset-II is that some sellers have less than 10 images. In an n-fold cross-validation (Kohavi,1995), we partition the images of seller's into n roughly equal subsets and repeat the training/testing process n times. In each round, one subset is treated as testing data while the other nine subsets are used as training data. In the training phase, we first use our proposed edge-based clustering algorithm to group the images of each seller's into one or more image groups in order to automatically discover a seller's editing styles. By applying the editing style summarization methods on image groups, each group is associated with an edge template, a mean color model, and a color probability model. In the testing phase, the proposed classification methods, i.e., edge-based classification, color-based classification, and color probability based classification, are applied to match a test image with an edge template, a color model, and a color probability model, or none at all (for images without editing styles). The authorship of a test image is then assigned to be the seller who owns the model/template that best matches that image.

Two evaluation scenarios, i.e., within a seller and across different sellers, are defined and used to measure the performance of the proposed framework.

1. *Within a seller*: In this scenario, we mainly focus on finding whether an image belongs to a specific seller within the scope of that seller. This kind of classification can be put into efficient use for account taken over detection when a fraudster tries to upload a product image that does not conform to the editing styles of that seller. In this scenario, if the test image is matched with any of the templates of that seller, that image is said to belong to that seller, and vice versa. Therefore, an image with some editing style is said to be correctly predicted if it is assigned the same seller's ID as indicated by its ground-truth label. Otherwise, the prediction fails. For an image without any significant editing style, a successful prediction is that the image is assigned a "no template" label after classification.

2. *Across different sellers*: In this scenario, we like to know whether the proposed framework can differentiate between editing styles created by different sellers. In this scenario, all the test images from each seller will be tested against all templates from all sellers. Since different seller accounts may have similar editing styles (e.g., multiple seller accounts associated with the same seller), therefore, if an image with some editing style is assigned to any template that summarizes that editing style, we say it is a correct predication. Otherwise, the prediction fails.

As aforementioned, there is a cutoff value used in each classification method proposed. In our experiments, we experiment with different cutoff values and determine experimentally a threshold value for each approach.

Figure 2 and Figure 3 present the experimental results for Dataset-I and Dataset-II, respectively. In Figures 2 and 3, the first and the second rows show the classification performance in "within seller" and "across sellers" scenarios, respectively.

In Figure 2, the best average accuracy values of the edge-classification method are 0.803 and 0.875 for "within seller" and "across sellers", respectively. The best average accuracy values of the color based classification are 0.843 and 0.799 for "within seller" and "across sellers", respectively. The best average accuracy values of the probability model classification are 0.535 and 0.871 for "within seller" and "across sellers", respectively.

In Figure 3, the highest average accuracy values of the edge-based classification method are 0.868 and 0.717 for "within seller" and "across sellers", respectively. The best average accuracy values of the color-based classification are 0.618 and 0.735 for "within seller" and "across sellers", respectively. And the best average accuracy values of color probability classification are 0.748 and 0.56 for "within seller" and "across sellers", respectively.

From the results, the proposed framework performs well on Dataset-I. However, there are cases where images with similar edge features on their product areas are grouped together during clustering, which may introduce noise into the summarization process. This indicates that removing a fixed percent of central area from each image cannot eliminate the product features entirely, causing false positives during clustering and classification.

In addition, we observe that the edge-based approach and the color-based approaches can complement each other. Our experimental results show that when we combine edge-based approach with either color-based summarization approach, i.e., the color-based approach or the color probability approach, the accuracy of classification can be drastically improved. For Dataset-I, the overall accuracies are 94% (edge + color-based) and 89% (edge + color probability model) for "within seller" evaluation.

We further analyze the result of each fold for Dataset-I and show the 10-fold accuracy values in Table I. The results in Table I show that the highest accuracy values of edge-classification method, across 10-fold cross validations, are 0.870 and 0.902 for "within seller" and "across sellers", respectively. For color-based classification, the highest accuracy values are 0.891 and 0.846 for "within seller" and "cross seller", respectively. And the corresponding accuracy values for color probability method are 0.804 and 1, respectively.

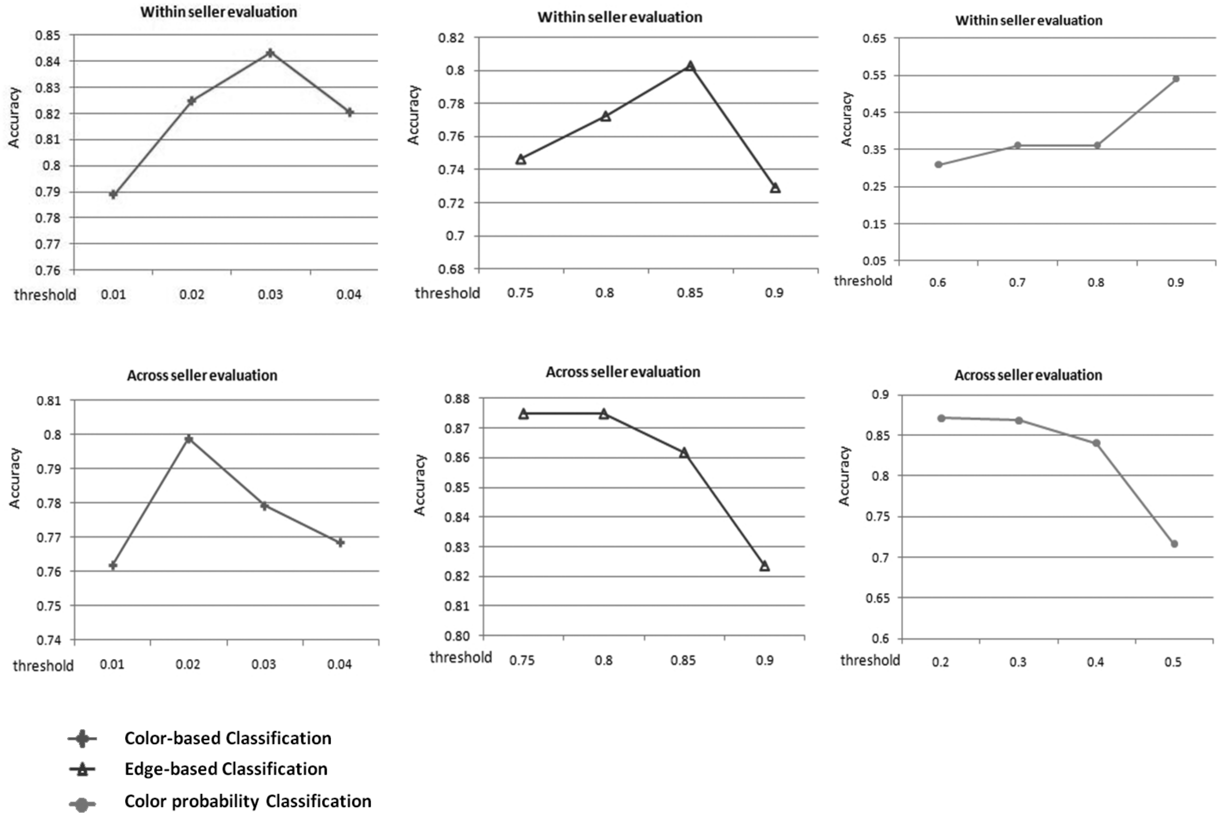

Figure 2. The evaluation results for images of 10 *sellers. The three columns from left to right are the results corresponding to the three classification methods, i.e., the color-based classification, the edge-based classification, and the color probability classification.*

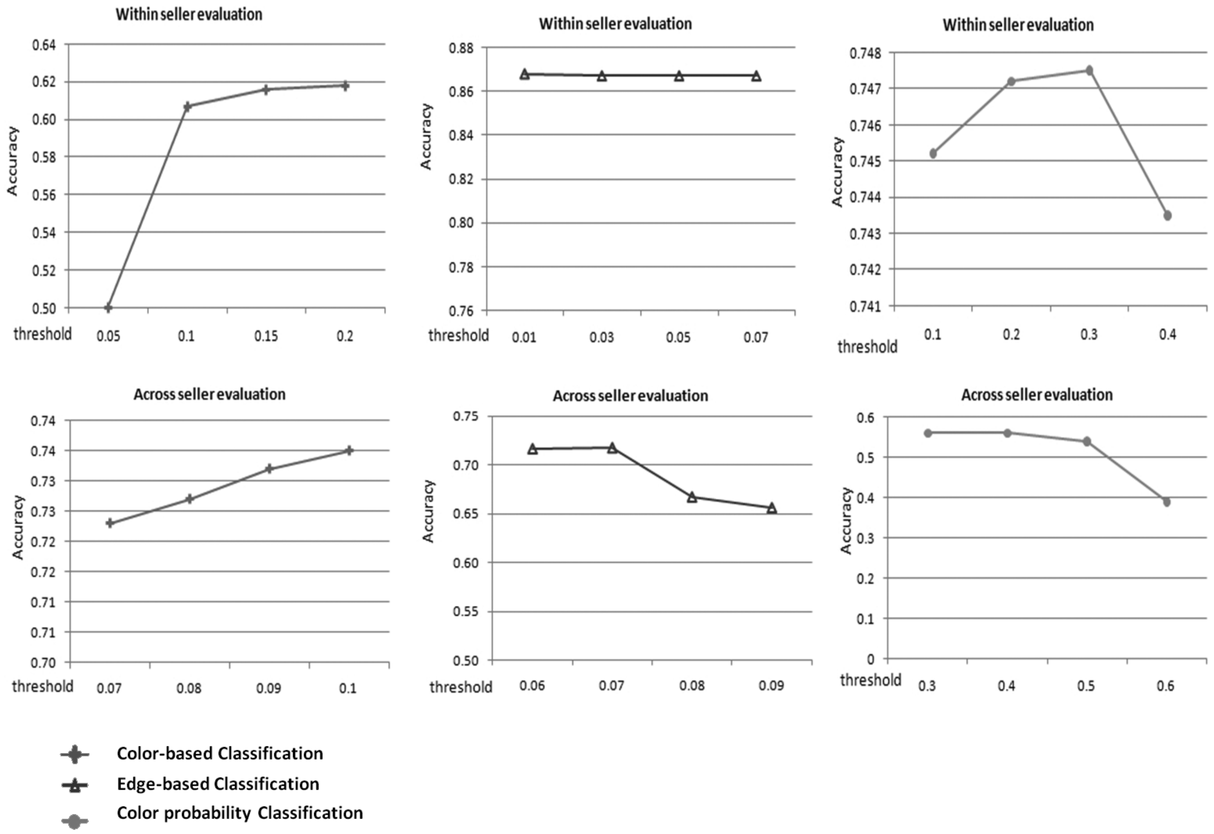

Figure 3.The evaluation results for images of 47 *sellers. The three columns from left to right are the results corresponding to the three classification methods, i.e., the color-based classification, the edge-based classification, and the color probability classification.*

From the "across sellers" evaluation results, we can observe that all the accuracies of the edge-based method are close to its maximum accuracy 0.9. However, for the color probability method, a larger range of accuracy values can be observed with the maximum accuracy being 1 and the minimum value being 0.370. For the "within seller" evaluation, the accuracies of the probability method are lower than that of the other two methods in almost all cases. The reason is that the probability method uses mean and standard deviation to represent the color distribution of training images. The standard deviation is largely affected by the size of the training set. In case a training set contains only 2~3 images, the color variance within the same group may still be high. As a consequence, dissimilar image could be incorrectly matched to the model. The performance of the color-based classification method is in between the other two methods and is the most time efficient.

Table 1. The accuracies of 10*-fold Evaluation for Dataset-I*

	Edge-based		***Color-based***		***Color Probability***	
i^{th} fold	***across***	***within***	***across***	***within***	***across***	***within***
1	0.835	0.835	0.813	0.846	0.703	0.176
2	0.882	0.763	0.839	0.817	1	0.613
3	0.860	0.828	0.774	0.807	1	0.462
4	0.891	0.761	0.750	0.870	1	0.011
5	0.859	0.837	0.837	0.870	0.637	0.294
6	0.902	0.837	0.739	0.804	0.370	0.489
7	0.880	0.870	0.794	0.891	1	0.283
8	0.870	0.739	0.826	0.837	1	0.804
9	0.901	0.769	0.846	0.868	1	0.022
10	0.868	0.791	0.769	0.824	1	0.154
max	0.902	0.870	0.846	0.891	1	0.804
min	0.835	0.739	0.774	0.807	0.370	0.011
mean	0.875	0.803	0.799	0.843	0.871	0.362

MULTI-ACCOUNT LINKING DETECTION

In the previous section, we identify the authorships of images based on their editing styles, and the authorship of each image is described by its assigned sellers' ID. The encoded editing styles (templates/models) can also be used to reveal the same owner behind multiple eBay accounts. We propose two methods to identify the linking of multiple accounts based on their editing styles. The first method, the covariance method, is based on the mean color model and uses covariance as a similarity measure to find similar mean color models. The second method, a clustering based method, uses Hough transform to group similar edge templates into groups. Since the mean color based model outperforms the color probability model in both datasets and is more time efficient than the latter, we only test the performance of the mean color based models in multi-account linking experiments, in addition to testing the edge-based models.

Similarity Measure by Using Covariance

Covariance (Baker, 1973) is a measure of how two variables change together. In this method, we define a similarity measure for mean color based models based on the covariance concept. More specifically, we calculate the similarity between two color models in each color channel separately. Then the sum of the similarities from all three channels is used as the similarity between the two models. To calculate the similarity in each channel, we sequence the pixels of each mean color model into one vector, and then calculate the covariance of the two vectors. The pixel values in each color channel in HSV color space are used as features of models in calculating the similarity.

In each channel in HSV color space, a color model is represented by a feature vector. t_i and t_j represent the feature vectors of two color models, respectively. We define the similarity between models t_i and t_j using the following correlation as defined in Eq. 4.

$$sim(t_i, t_j) = \frac{\mathrm{cov}(t_i, t_j)}{\sqrt{\mathrm{cov}(t_i, t_i)}\sqrt{\mathrm{cov}(t_j, t_j)}} \qquad (4)$$

where cov(t_i, t_j) is given by

$$\mathrm{cov}(t_i, t_j) = E[(t_i - E[t_i])(t_j - E[t_j])^T] \qquad (5)$$

We define a similarity matrix for the color models. After sorting the similarities, we start from the best matched model pair, which is a pair of models with the highest similarity value. Then, we find the other models that are matched with the pair of models, that is, if a model is sufficiently similar to either model in the best matched pair, that model will be linked with the model pair.

The steps to find similar color models are presented as follows:

1. Compute a similarity matrix *S* in which an element s_{ij} is the similarity of models t_i and t_j. Initialize $k = 0$ and choose a threshold value *th*;
2. Group[*k*] = [ ];
3. Find the best matched model pair s_{mn} whose similarity is equal to max(s_{ij}), set Group[*k*] = [t_m, t_n];
4. In the m^{th} and n^{th} rows of matrix *S*, find all the models whose similarity with *m* or *n* is greater than the threshold, then add those models into Group[*k*] and remove their corresponding columns and rows in *S*;
5. $k = k + 1$; go to Step 2.

In selecting the threshold value, we normalize the similarity scores between models to a value between 0 and 1. The threshold value is thus experimentally set to 0.95, roughly corresponding to 95% similarity.

Similarity Measure by Using Hough Transform

In this method, we apply Hough transform to cluster similar edge templates into groups. Similar to the proposed edge-based clustering method, this method uses Hough transform to find the best matched point of each pair of edge templates. Then image overlay is performed based on the best matched point to find the common area of the two edge templates. The similarity of two edge template is then computed according to Eq. 1.

A threshold value is selected so that if the similarity is greater than the threshold value, the two corresponding edge templates will be linked together. The threshold value is again experimentally set to 0.95 in this study. It means that if two edge templates are 95% similar with each other, they will be linked together. We start from an edge template, find all the templates linked to it based on the similarity measure and the threshold, and remove this group (cluster) from the subsequent clustering process.

Experiments

To evaluate the effectiveness of the proposed multi-account linking detection algorithms, we collected a larger dataset (Dataset-III) consisting of 6053 images from 100 seller accounts. In this dataset, half of the sellers do not use any editing style in their images while the other half of the sellers create at least one editing style in their images. Compared with the classification experiments, here the ground truth is defined in a different way which uses the actual templates/models representing the editing style of an image as the ground truth of that image rather than associating it with a specific seller's account. In this ground truth, images with the same editing style should be in the same group no matter to which seller account they belong. In collecting ground truth, we first visually identify all the distinct image editing styles in the dataset, and then assign the same label to images with the same editing style. We performed a 3-fold cross validation on Dataset-III, and the three sets of clustering results as shown in Tables 2 and 3.

To compare the resultant clusters with the ground truth, we use V-measure (Rosenberg & Hirschberg, 2007), a weighted harmonic mean of homogeneity (*hm*) and completeness (*cm*). V-measure is a conditional entropy-based method to evaluate the clustering results and is independent of the clustering algorithm being used. The definition of V-measure is given in Eq. 6, where β is a constant, which if greater than 1 would mean that cm is weighted β times more strongly than *hm*; otherwise *hm* is weighted more in the calculation. In this study, we compare our clustering results with the ground truth using this measure with different β value (ranging from 1 to 3). Tables 2 and 3 show the evaluation results.

$$v_\beta = \frac{(1+\beta^2)\times hm \times cm}{(\beta^2 \times hm) + cm} \qquad (6)$$

Table 2. V-measures of Color Model based linking

i^{th} **fold**	***hm***	***cm***	**v1**	**v2**	**v3**
1	0.73	0.71	0.72	0.71	0.71
2	0.66	0.70	0.68	0.69	0.69
3	0.65	0.70	0.67	0.69	0.69

Table 3. V-measures of Edge Template based clustering

i^{th} **fold**	***hm***	***cm***	**v1**	**v2**	**v3**
1	0.81	0.71	0.76	0.73	0.72
2	0.82	0.69	0.75	0.71	0.70
3	0.83	0.68	0.75	0.71	0.70

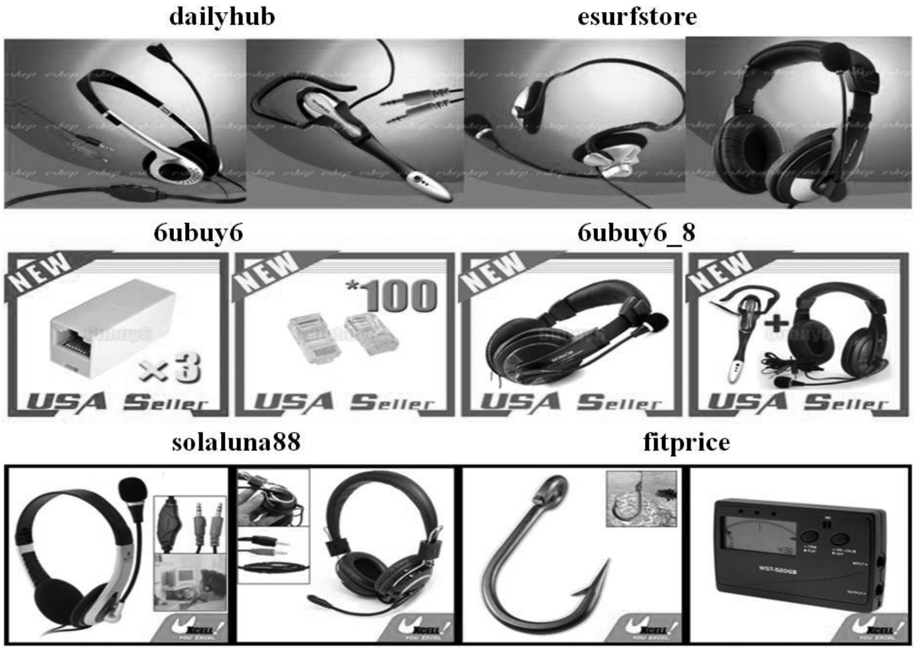

Figure 4. Examples of multi-account linking results: Each row shows the IDs of the sellers linked together and their sample images

We can observe that edge-based templates perform better than color based models, but as aforementioned, the former has a higher complexity than the latter. In Dataset-III, we also observe that there are three pairs of sellers that have almost identical editing styles. The three pairs of sellers are presented in Figure 4. Our proposed two algorithms successfully group each pair into the same group according to the similar editing styles present in each pair.

CONCLUSIONS

A framework of detecting sellers' authorship of eBay images is proposed in this paper by analyzing and encoding their distinctive image editing styles. Through a collaborative effort between the authors' institution and eBay, the output of this study will be used as added clues in eBay's seller profiling system to detect and prevent account taken-over and other related fraudulent behaviors. This framework includes three modules that target editing style summarization, classification and multi-account linking detection respectively. Three different summarization methods are developed to encode editing styles into templates (models) based on edge or color features, including an edge-based summarization and two color-based summarization methods. Accordingly, in order to predict the authorship of previously unlabeled images, we implement three classification methods corresponding to the three summarization methods, namely edge-based classification, color based classification, and color probability based classification. We applied these three algorithms to two image data sets that consist of product listing images downloaded from 10 sellers and another 47 sellers at eBay, respectively. *n*-fold cross-validation was performed to evaluate the experimental results. The three summarization and classification methods are compared, and our results indicate that the edge-based classification method is the most promising in identifying authorship of eBay images. The encoded image editing styles can also be used to predict the same owner behind multiple eBay seller accounts. In this study, we collected a larger data set consisting of 100 seller accounts. To measure similarity between seller accounts in terms of common image editing styles, we propose two methods for this purpose, including a covariance based method which uses covariance as a similarity measure for HSV color models, and a clustering based method by using Hough transform for edge template matching.

ACKNOWLEDGMENT

This work is supported in part by eBay Inc. and Dr. Chengcui Zhang's NSF grant (DBI-0649894).

REFERENCES

Love, H. (2002). *Attributing Authorship: An Introduction*. (pp.99-100). Cambridge: Cambridge University Press.

Abdel-Mottaleb, M. (2000). Image retrieval based on edge representation. *In Proceedings of 2000 International Conference on Image Processing*, 3, 734-737. Piscatway, NJ.

Fernandes, L.A.F., & Oliveira, M.M. (2008). Real-time Line Detection Through an Improved Hough Transform Voting Scheme. *Pattern Recognition*, 41(1), 299-314.

Zhao, M., Bu, J.J, & Chen, C. (2002). Robust Background Subtraction in HSV Color Space. *In Proceeding of SPIE: Vol. 4861. Multimedia Systems and Applications* (pp. 325–332). Boston, USA.

Zhao, Y., & Zobel, J. (2005). *Effective and Scalable Authorship Attribution Using Function Words*. Melbourne, Australia: RMIT University.

Robine, M., Hanna, P., Ferraro, P., & Allali, J. (2007). Adaptation of String Matching Algorithms for Identification of Near-Duplicate Music Documents. *In Proceedings of the International SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN)* (pp.37-43). Amsterdam, Netherlands.

Rafailidis, D., Nanopoulos, A., & Manolopoulos,Y. (2010). Building Tag-Aware Groups for Music High-Order Ranking and Topic Discovery. *Journal of Multimedia Data Engineering and Management (IJMDEM)*, 1(3), 1-18.

Hirose, S., Yoshimura, M., Hachimura, K., & Akama, R. (2008). Authorship Identification of Ukiyoe by Using Rakkan Image. *In Proceedings of the Eighth IAPR Workshop on Document Analysis Systems* (pp.143-150). Nara, Japan.

Jahne, B., Scharr, H., & Korkel, S. (1999). *Principles of filter design, Handbook of Computer Vision and Applications* (pp.125-152). Academic Press.

Baker, C. R. (1973). Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society,* 2, 273-289.

Sprinthall, R. C. (2002). *Basic Statistical Analysis, 7th ed.* (pp.43-58). Allyn and Bacon Publishers.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *In proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 2(12), 1137-1143. Morgan Kaufmann, San Mateo.

Rosenberg, A., & Hirschberg, J. (2007). V-Measure: A Conditional Entropy-based External Cluster Evaluation Measure. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)* (pp.410-420). Prague, Czech Republic.

Huang, C.H., & Wu, J.L. (2004). Attacking Visible Watermarking Schemes. *IEEE Transactions on Multimedia*, 6(1), 16-30.

Yu, D., Sattar, F. & Ma, K.K. (2002). Watermark Detection and Extraction Using Independent Component Analysis Method. *Journal on Applied Signal Processing*, 1, 92-104.

Pei, S.C., & Zeng, Y.C. (2006). A Novel Image Recovery Algorithm for Visible Watermarked Images. *IEEE Transactions on Information Forensics and Security*, 1(4), 543-550.

Chen, S.C. (2010). Multimedia Databases and Data Management: A Survey. *Journal of Multimedia Data Engineering and Management (IJMDEM)*, 1(1), 1-11.

Al-Asmari, A.K., & Al-Enizi, F.A. (2009). A Pyramid-Based Watermarking Technique for Digital Color Images Copyright Protection. *2009 International Conference on Computing, Engineering and Information* (pp. 44-47). Fullerton, California.

Lee, C.F., & Lee, H. E. (2008). A Blind Associative Watermark Detection Scheme Using Self-Embedding Technique. *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (pp.1122-1125). Harbin, China.