Annotating Historical Archives of Images

Xiaoyue Wang Lexiang Ye Eamonn Keogh Christian Shelton
Department of Computer Science and Engineering
University of California Riverside
Riverside, California
{xwang,lexiangy,eamonn,cshelton}@cs.ucr.edu

ABSTRACT
Recent initiatives like the Million Book Project and Google Print Library Project have already archived several million books in digital format, and within a few years a significant fraction of world's books will be online. While the majority of the data will naturally be text, there will also be tens of millions of pages of images. Many of these images will defy automation annotation for the foreseeable future, but a considerable fraction of the images may be amiable to automatic annotation by algorithms that can link the historical image with a modern contemporary, with its attendant metatags. In order to perform this linking we must have a suitable distance measure which appropriately combines the relevant features of shape, color, texture and text. However the best combination of these features will vary from application to application and even from one manuscript to another. In this work we propose a simple technique to learn the distance measure by perturbing the training set in a principled way. We show the utility of our ideas on archives of manuscripts containing images from natural history and cultural artifacts.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Retrieval models

Keywords
Historical Digital Libraries, Historical Manuscripts, Image Annotation, Information Extraction

1. INTRODUCTION
Recent initiatives like the Million Book Project and Google Print Library Project have already archived several million books in digital format, and within a few years a significant fraction of world's books will be online [10]. As Kevin Kelly recently noted, "*the real magic will come in the second act, as each word in each book is cross-linked, clustered, cited, extracted, indexed, analyzed, annotated, remixed, reassembled and woven deeper into the culture than ever before*" [12]. While this quotation explicitly singles out text, a similar argument can be made for *images*. Clearly the majority of the data gleaned from scanned books will be text, but there will also be tens of millions of pages of images. Many of these images will defy automation annotation for the foreseeable future, however a considerable fraction of the images may be amiable to automatic annotation by algorithms that can link the historical image with a modern contemporary, with its attendant meta tags [2]. As a concrete example, consider Figure 1.

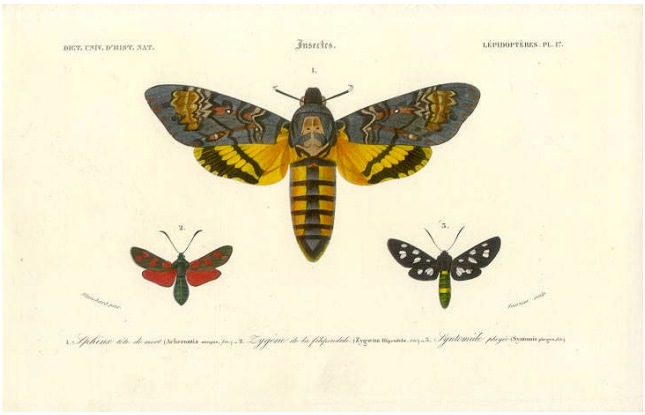

Figure 1: A page from a scientific text published in 1849 [6]. The heavily stylized scrip is difficult to read even at full resolution, however we have independently confirmed that the three insects are (left to right) *Zygaena filipendulae*, *Acherontia atropos* and *Syntomis phegea*

In this image the text annotations will surely defy even the state-of-the-art handwriting recognizers [1], and humans, particularly those without experience in reading cursive script are also unlikely to be able to parse these words. Suppose that we segment out the individual insects and search for the most similar images on the web (for the moment, we will gloss over the technical details of how this is done). In fact we have done this, and discovered the image in Figure 2.

Figure 2: An image of *Acherontia atropos*, also known as the Death's-head Hawkmoth, retrieved from URL [24]

The image is uncannily like the query image, we can confidently assume it is the same (or closely related) species and therefore we can link the historical image to its modern counterpart to provide context and annotations to the digital archive. In this example the shape and color provided the necessary clues to the identity of unknown object. More generally, different sets of features may be useful depending on the application. For example, most Diatoms (eukaryotic algae) are colorless when viewed at the microscopic scale, but are often richly textured, as in Figure 3.

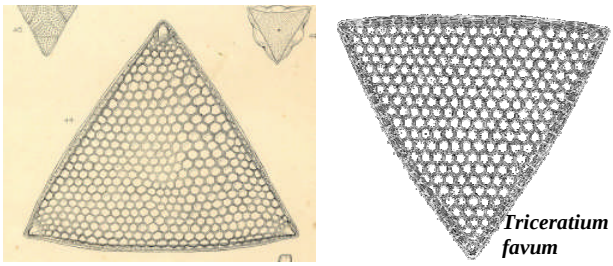

Figure 3: *Left*) Plate 5, fig. 44 of a Smith's *British Diatomaceae* (1853) [19], *Right*) An image of *Triceratium favum*, a type of Algae, retrieved from URL [25]

Here (rotation invariant) shape and texture allow us to link this object to a modern image, and learn of the latest research on this intriguing life form.

In this work we propose a general framework for annotating large archives of historical image manuscripts. Our work is similar in spirit to the work of Agosti et al. [2] on the automatic discovery of relationships among images in illuminated manuscripts; however we are focusing on the lower level primitives to support such work. We use different feature spaces such as shape, color and texture. We then we combine these similarities using appropriate weights. Our experiments show that the accuracy we can obtain is higher using a combined feature similarity measure than we can obtain using any individual single feature calculation. Our fundamental contribution is introducing a novel technique for learning this weighting parameter, in spite of a lack of any labeled training data.

The rest of this paper is organized as follows. In Section 2 we consider the necessary background and related work in image matching in the context of historical archives. In Section 3 we introduce our novel algorithm for learning the appropriate weighting parameter for combining different image features. Section 4 contains an empirical evaluation on several datasets which are up to five hundred years old. We concluded in Section 5 with a discussion of our results and directions for future work.

2. BACKGROUND AND RELATED WORK

The literature on image matching is vast, we refer the reader to [22] for an overview. Most of the work is concerned with efficiently and effectively matching images using one type of feature, i.e. shape, color, texture *or* text annotations. If we are to use more than one type of feature, we have the problem of finding an appropriate weighting parameter *w*. Research on combining two or more features tends to either assume that labeled training data is available [21][20], or it considers specialized domains where the value of *w* can be determined once and fixed forever. However it is clear that for the general problem of manuscript annotation the best value for *w* is highly data dependent. At one extreme, we may have monochrome engravings of objects as in Figure 4.*Left*, in which case we would wish to place all the weight on the shape features. For the other extreme, imagine we are matching heraldic shields as in Figure 4.*Right*, here there is very little variation in shape (and none of it meaningful), and we would wish the algorithm to consider color only[1].

[1] This is true in our particular dataset, see Figure 5 and Figure 17. However in other datasets of heraldic shields the shapes *can* be very useful.

Figure 4: Two examples at the extremes of color/shape importance for matching. *Left*) Historia Naturalis (1648) by John Johnston (engraved by Matthaeus Merian). *Right*) Coats of arms. fols. 25v-26r. Founders' and benefactors' book of Tewkesbury Abbey. Early 16th century.

There are many existing techniques for learning this mixing parameter *w*, if we have access to subjective similarity judgments [20][16][21]. While we would not rule out human interaction to *refine* a distance measure in an important domain, the scale of the problems we wish to eventually consider means that we would like to have a completely automated system to at least bootstrap the process and produce an initial high quality measure.

There are also dozens of methods for learning distance measures if we have *labeled* training data. One basic idea is to use a wrapper over all possible feature weights [4], another idea is to set up a classification problem where the input is two objects and the output is 0 if they are in the same class and 1 if they are not. We can then train a classifier which provides continuous output, like an SVM or neural network [11]. The continuous output of the trained classifier can then be used as a distance measure [14]. The problem with all these approaches is that they require labeled training data. However we are explicitly assuming that we do not have any such labels. We simply have a collection of objects manually or automatically extracted from documents. As we shall see in Section 3, our solution is to produce pseudo-labeled data and use it to learn the *w* parameter.

2.1 Image Matching Primitives

While we plan to present a generic technique to allow us to find a mixing parameter *w* for any of the many measures defined for shape, color, texture or text features, for concreteness will we will show the particular shape and color measures we use in the diverse experiments in this work.

2.1.1 Color Matching

While many information retrieval uses of color use a single color histogram to represent the entire query object, it is clear that in many areas of historical manuscript annotation we need to consider the localized arrangements of color. Consider for example the four heraldic shields shown in Figure 5. It is clear that all four must have near identical color histograms, yet they are clearly distinct.

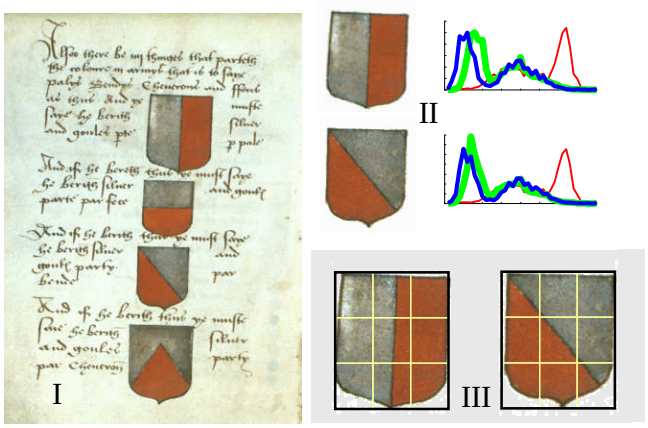

Figure 5: I) Leaf 36v from *Treatises on Heraldry*, dating to the 15th century. II) The shields have virtually identical color histograms. III) By creating localized color regions we can easily distinguish between the objects.

While this example is particularly clear and obvious, we have also observed similar cases for butterflies and other natural objects. Our solution is to localize the color information by creating a grid within the Minimum Bounding Rectangle (MBR) of the object, and considering the color histogram for each cell. For simplicity, and to help mitigate the problem of overfitting, we create an equal number, g, of row and columns. This leaves open the question of how we set the value of g. We propose a simple method to do this. The intuition of our idea is that if g is too small, there will be a lot of variance of color within a cell, but when each cell contains a single patch of color, the variance within each cell should be low. We can therefore search over different values of g and measure the change in average variance within the g^2 cells as we increase g. More formally:

$$g = \text{argmax}\{\ \text{avg}(\text{var}(g_i)/\text{avg}(\text{var}(g_{i+1})\ \},\ 0 \le i,\ g_0 \equiv g_1$$

This is rather similar to the criteria of *information gain* used in decision trees. In Figure 6, we tested this idea on two small contrived datasets for which the objectively correct answer is obvious.

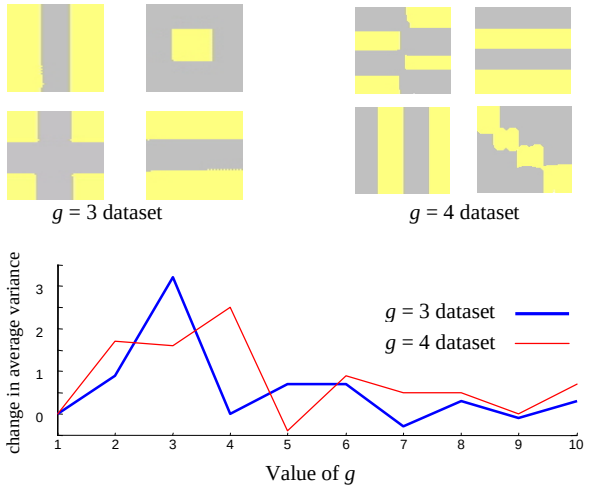

Figure 6: *Top*) Two synthetic datasets for which the best value of g is known. *Bottom*) For both datasets, our heuristic picks the correct value for g.

In Figure 7 we test the heuristic on a real dataset for which, subjectively speaking, there is a narrow range of reasonable choices for g.

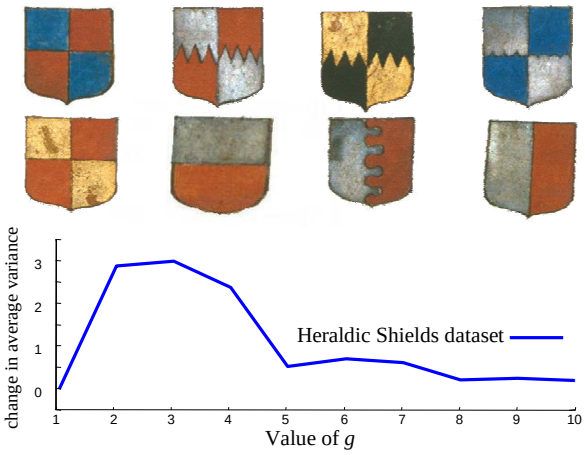

Figure 7: *Top*) A subset of heraldic shields from *Treatises on Heraldry*, which subjectively seem to require a value of g which is 2 or 3. *Bottom*) The value of g chosen by our heuristic seems plausible for this dataset.

In this example the heuristic gives a plausible answer, as it does in all datasets considered in this paper. We therefore use this idea in all experiments in this work.

2.1.2 *Shape Matching*

There are literally hundreds of shape measures in the literature; [22] and the references therein provide an excellent overview. In choosing a shape measure for mining historical archives we have two critical requirements. We must have a scalable algorithm, given that we may need to compare millions of images, and we must have shape measure that requires few parameters. Some shape measures require as many as 8 parameters. While it may be possible to tune these in limited domains for which we have massive amounts of labeled data, for the task at hand we need to compare unlabeled historical images to unconstrained images retrieved from the web. Any attempt to tune parameters is very likely to result in over fitting. As noted above, we will eventually need to find a parameter w to combine the contribution of shape and color. Having additional parameters for just the shape measure will increase the search space, resulting in slower training times and dramatically increasing the possibility of over fitting.

Fortunately there is at least one shape distance measure which is completely parameter-free and scalable to large datasets [13]. The basic idea to transform the two-dimensional shape to a one-dimensional "time series". Figure 8 gives a visual intuition has to how this is achieved. The distance from every point on the profile to the center is measured and treated as the Y-axis of a time series of length n.

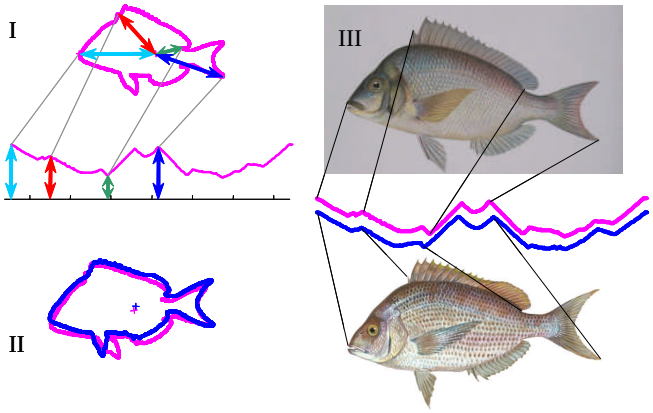

Figure 8: A visual intuition of the conversion of a two-dimensional shape to a one-dimensional "time series". II) Two shapes that are similar in the shape space will also be similar in the time series shape. III) Here we compare an 1890 chromolithograph [5] to a modern photograph of *Stenotomus chrysops* (common name: Scup or Porgy)

Once we have transformed the shape into a time series, we can leverage off the wealth of distance measures and indexing techniques available for time series [13]. In this work we use Euclidean distance as the distance measure. Recent work has shown that the Euclidean distance on this representation is at least competitive with more complex measures on a large variety of shape matching problems [13]. Note that in the example in Figure 8 the two fish are pointing in the same direction and have approximately the same rotation. However we can achieve invariance to both enantiomorphic and rotated shape with a very small overhead [13].

2.1.3 Texture and Text Matching

For simplicity and in order to tightly focus our experiments, we only consider shape and color for the rest of this work. We note however that our ideas can generalize to other features that can be measured from the historical documents. One obvious feature is texture (recall Figure 3). Texture features may be useful in some domains where shape and color do not completely disambiguate the objects in question, for example consider the heraldic shields shown in Figure 9.

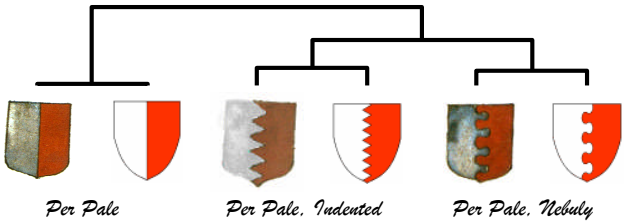

Figure 9: Heraldic shields from the 15th century *Treatises on Heraldry*, with three modern equivalents, clustered using Gabor filters and group average linkage.

Text may also be a very useful feature for the task at hand. However extracting text from historical documents can be very difficult. A recent special issue of the International Journal on Document Analysis and Recognition on the topic highlights the challenges and recent progress [3]. Like the other features of shape, color and texture, any similarity measures for text would need to allow matching under uncertainty, because the inevitable errors in optical character recognition are further compounded by the document quality of many historical archives and the potential for spelling changes over time [7]

3. LEARNING A DISTANCE MEASURE

For simplicity we are considering just the simple case where we want to find a good value for w, a weighting parameter to combine color and shape:

$$\text{Combined\_Dist}(a,b,w) = w * \text{Dist}_{\text{color}}(a,b) + (1- w)* \text{Dist}_{\text{shape}}(a,b)$$

$$0 \le w \le 1$$

Intuitively a good value is one that will maximize the number of correct mappings between the objects in our historical manuscript collection and the collection of real world images. We are making no assumptions about the collection of real world images. They may be highly constrained and structured, for example the set of idealized prototypes of heraldic shields shown in Figure 9, or they may be highly heterogeneous, for example the set of all images returned to a Google image query for "heraldic shields".

To make sure that the shape and color distance measures are commensurate we normalize them. Once we have the shape/color distance matrix between the drawing collection dataset and the real world image dataset, we first find the maximum entry in the matrix, and then we divide the whole matrix by this maximum value. To motivate our solution to the problem of selecting a good value for w, let us imagine an idealized case. Suppose our collection of historical images happens to consist of only labeled pairs, for example two Monarch butterflies, two Viceroy butterflies etc. If this were the case, we could determine a good value of w as follows. We could split the dataset into two subsets A and B, such that each of subset contains exactly one of the labeled pairs. We can then find the nearest neighbor for each object in A from the subset B while varying the value of w from zero to one. This is essentially 2-fold classification evaluation. A good value of w is the one that maximizes classification accuracy. While this idea is obvious and intuitive, for concreteness we outline the code in Table 1.

Table 1: An algorithm to learn the parameter *w*

Algorithm Learn_Weighting_Parameter(*A*,*B*)	
1	Calculate the normalized color/shape distance matrices Dis_{color} /Dis_{shape} between *A* and *B*;
2	**for** $w = 0$ **to** 1 **in steps** of 0.1
3	$Combined\_Dist = w * Dist_{color} + (1 - w) * Dist_{shape}$;
4	$accuracy(w) = 0$;
5	**for** each object *i* in *A*
6	find *i*'s nearest neighbor *j* in *B*;
7	**if** *i* and *j* have the same class label
8	$accuracy(w) = accuracy(w) + 1$;
9	**end**
10	**end**
11	**end**
12	Find the w_{max} with the greatest *accuracy*;
13	**return** w_{max};

The obvious problem with the scenario outlined above is that we do not in general have two of each object. In fact, we are assuming that the objects are not annotated, so we have no labels

of any kind. Nevertheless we can use this idea, by creating *synthetic* pairs of objects. The idea is that for each object in the set of historical manuscripts *A*, we create a new example of it and place these objects in set *B*, we can then use the algorithm in Table 1 directly.

The success of this idea relies upon our ability to produce realistic pairs. At one extreme, if we simply duplicated each object, there could be no "distortion" to learn, and any value of *w* would give perfect accuracy. At the other extreme, if we simply randomly create objects and label them in the same class then we should expect the classifier to perform at close to the default rate, since there is no structure to learn from.

We have created two strategies for creating the necessary pairs. The first is very simple, but only works for objects which are symmetric about at least one axis. While this is the case for many objects of interest, including butterflies/moths, bird eggs, arrowheads etc, it is not true for some natural objects, including mollusks shells, lateral views of most macroscopic animals and some cultural artifacts. Because of this we have also created a generic technique for creating synthetic pairs which does not require symmetry. We discuss both methods in the following two sections.

3.1 Exploiting Chirality[2]

The intuition here is to assume that there is one idealized object for each class. This idealized object can never be reproduced exactly, either by drawing it or photographing it. Any attempt to reproduce it must invariably produce "distortions", and it is the purpose of our weighted distance measure to be invariant to these distortions. We want to determine the weight that is most invariant to the distortion realized both in drawing and photographs. Although we cannot establish this in general, the "two-of-each" assumption above allows us to empirically determine the best weight for invariance to the distortion realized just drawings, and we hope that this weight will generalize to photographs too.

Note that many of the objects of interest in historical manuscripts are approximately symmetrical. This suggests we can attempt to create synthetic data simply by flipping existing images from left to right along the central axis. Although idealized butterflies are perfectly symmetrical, like most real world items, actual butteries, and drawings or photographs of them, are never perfectly symmetrical in either shape or color, as shown in Figure 10. We therefore do not get the exact same butterfly after reversal, and we can treat the flipped image as a new example.

Figure 10 : Examples flipped near-symmetrical objects taken from historical archives. *Left*) A *Per Bend* heraldic shield. *Center*) A *Danaus plexippus* butterfly. *Right*) A projectile point

In the examples in Figure 10, our algorithm would find that reversing a heraldic shield makes little difference to the shape, but can make a *large* difference to the color distribution. In contrast, reversing the projectile point makes no difference to the color distribution, but affects the shape (note the asymmetric tangs at the bottom of the shape). Finally for the butterfly example, the reversal makes subtle differences to both color and shape, suggesting (correctly, as we shall see) that both color and shape are important in this domain. Of course, here we are only looking at one example from each domain, there may be other examples where reversal does not have the desired effect, for example two of the heraldic shields in Figure 5 are near identical after reversal. However for the algorithm in Table 1 to work we only require that some fraction of the datasets offer clues to the relative importance of shape and color by comparing to their enantiomorphs.

3.2 Pseudo-example Generation

As we shall demonstrate in Section 4, the simple technique to exploit chirality proposed in the previous section works surprisingly well. However, we need to have a more general technique to create synthetic examples, given that we may have a dataset of intrinsically asymmetric objects. We propose to do this by averaging objects. Concretely:

For each object a_i in *A*, we create a new object b_i by averaging a_i's shape with the shape of its nearest neighbor considering only shape, and by averaging its color with the color of its nearest neighbor considering only color. The set of all newly created objects becomes the set *B*. Figure 11 gives a visual intuition of this process.

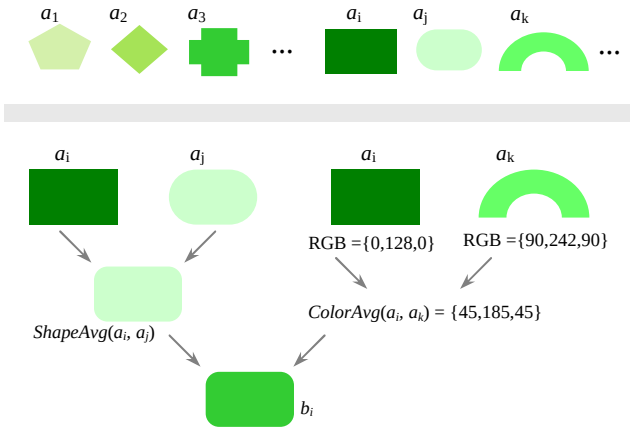

Figure 11 : A visual explanation of the process of producing a new synthetic example of the object a_i. *Top*) A set of shapes *A*. *Bottom*) A slightly distorted version of object a_i is created and labeled b_i. Object a_j is the nearest neighbor to a_i if we consider only shape, and a_k is the nearest neighbor to a_i considering only color

As the figure suggests, we can generally expect the newly created object b_i to be similar to its "parent" a_i without being identical. In this toy dataset, only the shape of the objects is meaningful; the colors are simply random shades of green. As we show in Figure 12, if we later attempt to classify object b_i by finding its nearest neighbor is set *A*, the value of the parameter *w* used in the Combined_Dist function is critical. It is this criticality that lets us automatically discover that shape is the key to this particular dataset.

[2] A figure is chiral (and said to have chirality) if it is not identical to its mirror image.

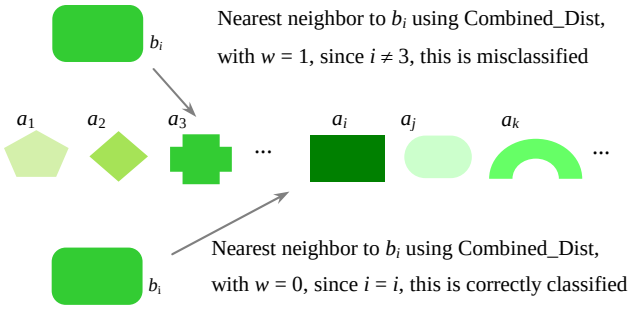

Figure 12 : If we attempt to classify b_j using the nearest neighbor algorithm and the Combined_Dist function, we find the setting of the parameter *w* is critical to success

Let *y* denote the class label of a particular example and let $x_1, \ldots, x_n$ denote the different features domain values of this example (a given x_i might have multiple components, but they all correspond to one measurement modality). We assume that each domain x_i lies in a space with a distance metric, but that the collection $X = (x_1, \ldots, x_n)$ does not have an a priori defined distance measure.

We are given a set of examples $\{X_1, \ldots, X_m\}$ and their associated class labels, $\{Y_1, \ldots, Y_m\}$, which collectively we will denote *D*. We would like to use this data to find a global distance measure on *X* as a weighted sum of the individual distance measures on its subcomponents:

$$d(X, X') = \sum_{i=1}^{n} w_i d(x_i, x_i').$$

There have been many prior methods for learning such weighted distances for nearest neighbor classifiers. However, all of these algorithms assume that each class label is represented more than once in the training set. In our application, each of the training set examples are of different classes. Stated differently, our goal in classification of a new testing instance is to find the training instance to which it is most similar, not merely to find the general class of training instances to which it is most similar.

Therefore, we cannot adjust the distance weights to achieve good leave-one-out training set accuracy. Our method of generating pseudo-examples on which to tune the distance measure weights is a natural consequence of the model described below.

Assumptions

Assume that the joint distribution over examples from which the training and testing set are drawn factors when conditioned on the label:

$$p(X, y) = p(y)\prod_{i=1}^{n} p(x_i \mid y). \qquad \textbf{(1)}$$

This is the same assumption made for naive-Bayes classification, except that we are only making the assumption at the level of the different features domains, not also within a domain. We further assume that each class label is equally likely ($p(y)$ is a constant). With only one example from each class, this is a natural assumption.

We can now condition this distribution on the distribution over the label of an example *X*.

$$p(y \mid X, D) = \frac{p(y)\prod_{i=1}^{n} p(x_i \mid y, D)}{p(X \mid D)} \qquad \textbf{(2)}$$

$$= \frac{p(y)}{p(X \mid D)}\prod_{i=1}^{n} p(x_{i|} y, x_i^{f(y)}) \qquad \textbf{(3)}$$

where $f(y)$ returns the index in the training set of the example with class label *y* (recall each label is unique in the training set). This last step is a result of the dependence assumption made in equation 1: the posterior distribution over the features in domain *i* for examples of class *y* depends only on the single example in the training set of class, and only on its features for domain *i*. We now make one final assumption:

$$p(x_i \mid y, x_i^{f(y)}) = g_i(d(x_i, x_i^{f(y)})) \qquad \textbf{(4)}$$

where g_i is some monotonically decreasing function. That is, for feature domain *i*, assume that the posterior distribution of examples from a class decreases monotonically with the distance from the single example we have of that class. Given that we are relying on the distance metric for feature domain *i*, this is a natural assumption.

Thus, given an example *X* with label *y*, we would like to construct a new example that we feel relatively certain would have the same label. Thus, the probability that *X* has label *y* should be greater than the probability it has any other label, *y*':

$$p(y \mid X, D) > p(y' \mid X, D) \qquad \forall y' \neq y$$

$$\prod_{i=1}^{n} g_i(d(x_i, x_i^{f(y)})) > \prod_{i=1}^{n} g_i(d(x_i, x_i^{f(y')})) \qquad \forall y' \neq y$$

To meet this condition it is sufficient (but not necessary) that

$$g_i(d(x_i, x_i^{f(y)})) > g_i(d(x_i, x_i^{f(y')})) \qquad \forall i, y' \neq y$$

$$d(x_i, x_i^{f(y)}) < d(x_i, x_i^{f(y')}) \qquad \forall i, y' \neq y$$

Thus, if every feature set for our new point is closer to the target example than to any other example, the new point is more most likely to have the same label as the target.

So, any example we generate by distorting the domains of the target toward its nearest neighbors will generate an example that is most likely to be an example of the target's class, provided that warping is less than half the distance between the two objects. Instead of selecting many such distortions, we choose to select the single most extreme valid distortion. If this single new example is correctly classified, then any example generated from a lesser distortion would also be correctly classified. Our theory states that we can select any distortion up to, but not including points exactly halfway between the target and its nearest neighbor. However, we have found no ill effects of employing a distortion of exactly 0.5, instead of the theoretical bound of $0.5 - \varepsilon$.

4. EXPERIMENTS

In this section, we conduct experiments to demonstrate the utility of our ideas. All experiments are designed to be completely reproducible, the datasets and additional details can be downloaded at [23]. We test both the exploiting chirality approach and the more general Pseudo Example Generation (PEG) method.

4.1 Pseudo-Example Generation of Insects

We extracted 84 hand-drawn insects from historical manuscripts either available online or from historical books including [15][18]. The ages of the images in question range from 1658 to 1964. In spite of their age, almost all of the images are colored, as it was common practice in centuries past to print books in black and white, and offer luxury editions that had been hand colored. Since the majority of insects considered are butterflies, we will denote this dataset "*butterflies*" for simplicity.

The extraction of the insects from the original archives was partly manual, however in a parallel work we are exploring (with initially promising results) completely automatic methods to extract objects of interest from arbitrary texts.

With the 84 drawn butterflies as dataset *A*, we obtain the dataset *B* using PEG, as described in Section 3.2. Figure 13 illustrates the classification accuracy for different values of *w* on the butterflies dataset.

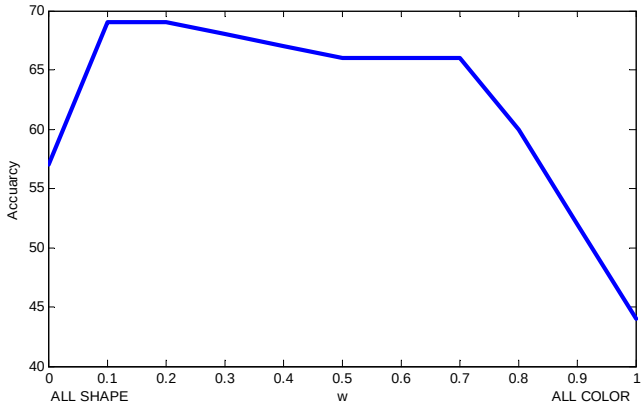

Figure 13: Classification accuracy on the butterfly dataset when the set *B* is obtained using PEG.

The classification accuracy is maximized when *w* equals 0.1 or 0.2. The result is similar to (but not exactly the same) as the result we gain from the experiment on the same dataset using the chirality method (as discussed below). The results suggest that we can benefit from a combination of shape and color, and that using *only* shape or *only* color would produce inferior results.

Having discovered the parameter setting, we can use the Join_by_Combined_Measure algorithm shown in Table 2 to link our 84 hand-drawn butterflies' dataset (*A*) with a larger reference dataset (*R*). In this case dataset (*R*) consist of 852 real insect images collected from various WWW sources, and the parameter w_{max} = 0.2 as learned above. We made sure that at least one of each of the 84 drawn butterflies appears in the larger collection, base on the insect's species. However we made no effort to make sure that the shape, rotation or colors are the same.

Table 2:

Algorithm Join_by_Combined_Measure(A,R, w_{max})	
1	Calculate the normalized color/shape distance matrices $Dist_{color}$ /$Dist_{shape}$ between A and R;
2	$Combined\_Dis = w_{max} * Dist_{color} + (1 - w_{max}) * Dist_{shape}$;
3	**for** each object i in A
4	find i's nearest neighbor j in R;
5	$pair(i) = j$;
6	**end**
7	**return** *pair*;

For all images used, we identified the species either from the source material, or we had entomologist Dr. Agenor Mafra-Neto identify them. With the help of Dr. Mafra-Neto we divided the results into 3 categories, perfect matches, not perfect but plausible matches and poor matches. The category "perfect but plausible" is awarded to matches which are not correct at the species level, but are in the same genus, or to insects known to be mimics of each other.

In this experiment we had 16 perfect matches, 21 plausible matches and 47 poor matches in total. Figure 14 shows some representative examples of matches.

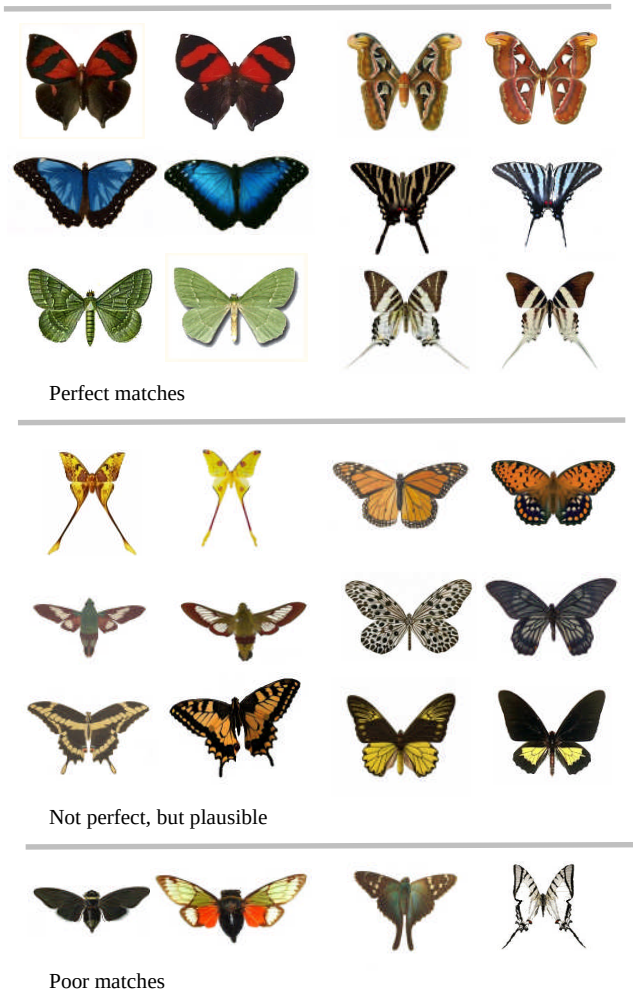

Figure 14: Sample matches from the butterfly experiment.

4.2 Exploiting Chirality of Insects

Given the symmetric nature of insects, we also considered our technique for exploiting the slight chirality apparent in drawings of insects.

In this experiment, the set *A* is the original dataset of 84 hand-drawn butterflies. Set *B* is obtained by mirroring all the images of *A* from left to right and explained in Section 3.2. We use the algorithm Learn_Weighting_Parameter shown in Table 1 to learn the parameter *w*. Figure 15 shows the classification accuracy as a function of *w* ranging from 0 to 1 with steps of 0.1. It shows that we have the greatest accuracy when w_{max} = 0.4, which appears to be significantly better than either of the shape measure only or the color measure only.

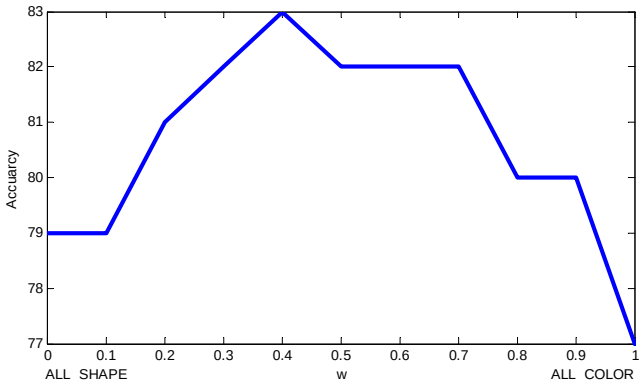

Figure 15: Classification accuracy on the butterfly dataset when the set *B* is obtained by exploiting chirality. When *w* = 0 all the weight is on the shape features and when *w* = 1 all the weight is on the color features.

After performing a join with the same 852 real insect images used in the previous section we had 15 perfect matches, 19 plausible matches and 50 poor matches in total. This is slightly worst that the PEG method, but still surprisingly good given the complexity of the problem.

4.3 Exploiting Chirality of Heraldic Shields

For the heraldic shields dataset we do not currently have a complete ground truth dataset. For our experiments, we created a large set random set of shields by taking 25 common shield patterns and randomly coloring them from a palette of 20 colors. To make the problem more realistic, we also added Gaussian noise with mean 0 and variance which is 1% of the original variance, to the images, and randomly distorted the shape of the templates by changing their height and width respectively by a random amount chosen uniformly in the range of -10 to +10%. In the end we have a dataset of 2,350 synthetic heraldic shields.

The historical archive consists of 100 images extracted from Treatises on Heraldry, dating to the 15th century. The original manuscript is housed in the Bodleian Library in Oxford.

Using the 100 hand-drawn images as the set *A* for the Algorithm Learn_Weighting_Parameter(*A*,*B*) shown in Table 1, we created the set *B* using the exploiting chirality Section 3.2. In Figure 16 we show the result of the experiment on this dataset.

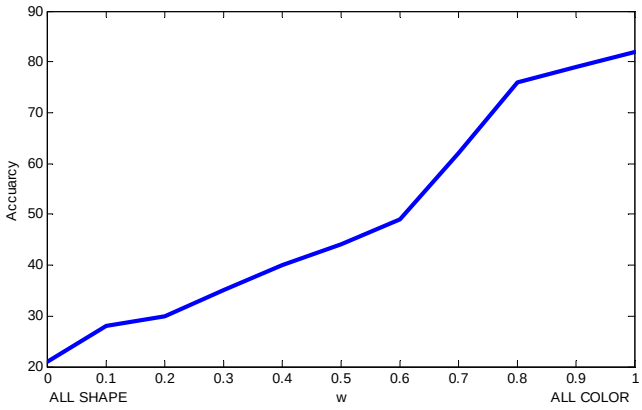

Figure 16: Classification accuracy on the heraldic shields dataset when the set *B* is obtained by PEG.

As we can see the classification accuracy improves when *w* increases from 0 to 1. Figure 16 shows that the classification accuracy is maximized when *w* = 1. The above result suggests that only color information is important in this particular heraldic shields dataset. This is because these objects have identical outlines (within the limits of the artist's ability) whereas their colors are very diverse. Therefore it is not surprising that we have greater accuracy when we employ only the color.

We used the algorithm Join_by_Combined_Measure shown in Table 2 to combine the color and shape features on heraldic shields dataset, where *X* is the 100 drawing heraldic shields images and *Y* is the 2,350 synthetic heraldic shields, and w_{max} = 1.

In the absence of ground truth we divided the results into 3 categories, perfect matches, not perfect but plausible matches and poor matches. Figure 17 shows examples of each type. In total, we had 16 perfect matches, 19 plausible matches and 65 poor matches. The vast majority of the poor matches are simply patterns which did not exist in our database, or where the hand-drawn historical image was highly degraded.

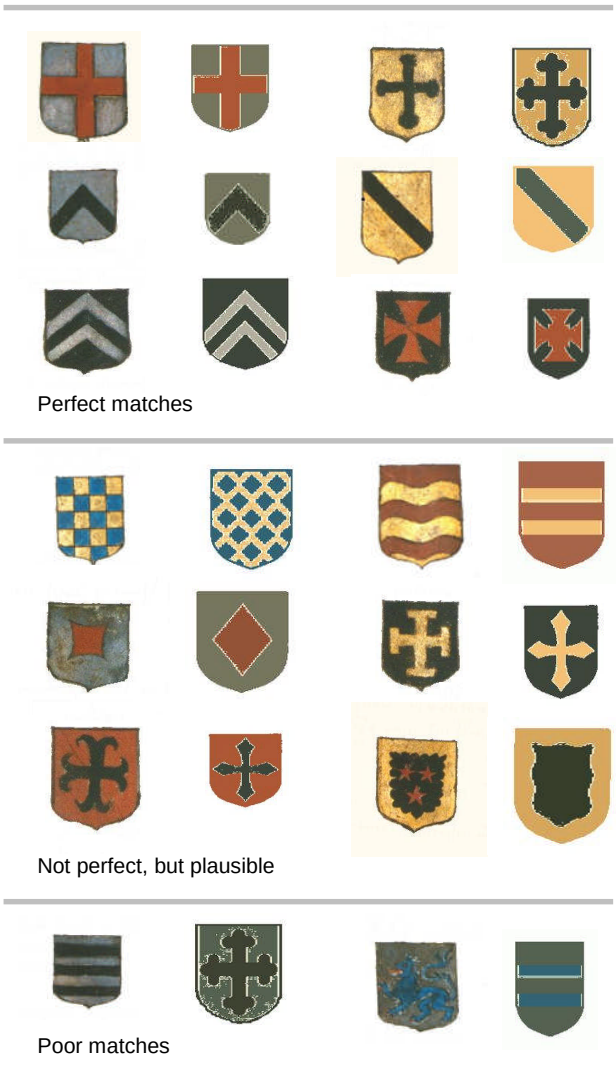

Figure 17: Sample matches from the Heraldic shield experiment.

Experiments using the PEG technique produced near identical results and are omitted due to space limitations.

4.4 Exploiting Chirality of Projectile Points

In the two previous examples we had a situation where we expected using just color to be best (heraldic shields), and a mixture of color and shape to be best (butterflies). For completeness we now consider a dataset where we strongly suspect that *only* shape matters. In particular we examined a set of drawings of projectile points (arrowheads) from a field archeologist's notebook [9]. While the sketches capture the shape (and to some degree the texture) of the objects in question, they are clearly devoid of any color information.

In this experiment, 30 images of hand-drawn arrowheads, taken from historical documents [9], are used as the set A for the algorithm Learn_Weighting_Parameter. As before, we produce the set B by reversing the 30 drawn arrowheads of set A from left to right. Figure 18 shows when w goes up above 0.5, the classification accuracy drops dramatically. The figure implies that the shape information is of greater importance than the color information in this application.

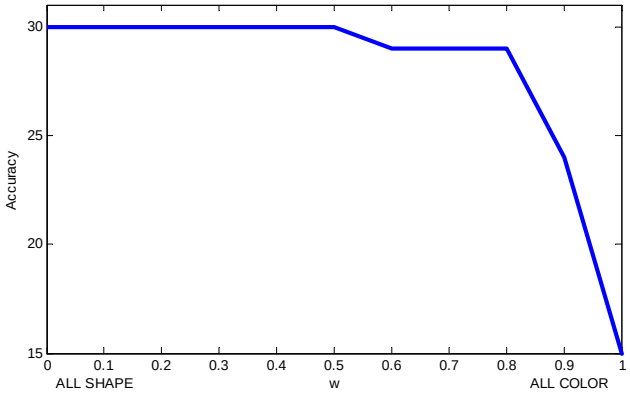

Figure 18: Classification accuracy on the arrowheads dataset when the set *B* is obtained by exploiting chirality.

Note that we have the problem here of breaking ties, since the accuracy is maximized over the range of $w = [0, 0.5]$. Ties may be common using our technique, given that the range of values for the classification accuracy is a relatively small integer (i.e. $|A|$).

We can break ties by choosing the value in the maximizing range of w that minimizes the sum of all distances to the nearest neighbors for *just* the correctly classified examples.

We linked our 30 images with a collection of 310 arrowheads from URL [26]. Figure 19 shows examples of the five best matches as measured by the Combined_Dist function.

Figure 19: Sample matches from the projectile point experiment. a) Zella-Grahm-Mentzer, b) Delhi-Perry, c) Firstview, d) Williams-t1525, e) Zorra-1802

Experiments using the PEG technique produced near identical results are omitted from brevity.

5. CONCLUSIONS

In this work we consider the problem of annotating images in historical archives, a problem sure to gain more attention as increasing numbers of books are digitally archived. We showed that a critical issue is determining the appropriate mix of shape/color/texture to consider, and we introduced a novel algorithm to determine this. Future work will consider efficiency issues, which will become more important as we attempt to scale our ideas to larger datasets.

6. REFERENCES

[1] T. Adamek, N. E. O'Connor, and A. F. Smeaton. Word matching using single closed contours for indexing handwritten historical documents. *IJDAR 9(2-4): 153-165 (2007)*.

[2] M. Agosti, N. Ferro, and N. Orio. Annotations as a Tool for Disclosing Hidden Relationships Between Illuminated Manuscripts. *AI*IA 2007: 662-673*.

[3] A. Antonacopoulos, and A. C. Downton. Special issue on the analysis of historical documents. *IJDAR 9(2-4): 75-77 (2007)*.

[4] X. Chen, and T-J. Cham. Learning Feature Distance Measures for Image Correspondences. *CVPR (2) 2005: 560-567*.

[5] S. F. Denton (1890). Report of the Fish and Game commission of The State of New York. Seventh Report.

[6] C. D'Orbigny (1849). Dictionnaire Universel d'Histoire Naturelle. *Renard & Martinet, Paris*.

[7] A. Ernst-Gerlach, N. Fuhr (2007) Retrieval in text collections with historic spelling using linguistic and spelling variants. JCDL 2007: 333-341

[8] P. S. Fres (1991). The Illustrated Encyclopedia of the Butterfly World, by, 1991. *Tiger Books International PLC, London or 1989, Crescent Books, Crown Publishers Inc, NY*.

[9] P. Gregory. (1958). Guide to the identification of certain American Indian projectile points. The Norman Society. *Special bulletin no. 4 of the Oklahoma Anthropological Society*.

[10] M. Herwig (2007). GOOGLE'S TOTAL LIBRARY: Putting the World's Books on the Web. *http://www.spiegel.de/international/*.

[11] T. Joachims. Making large-scale support vector machine learning practical. In *Advances in Kernel Methods: Support Vector Machines*. MIT Press, Cambridge, MA, 1998.

[12] K. Kelly (2006). Scan This Book! *N.Y. TIMES, May 14, § 6 (Magazine), at 42*.

[13] E. Keogh, L. Wei, X. Xi, S. H. Lee, and M. Vlachos. LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures. *In Proceedings of Very Large Databases (VLDB'06), 2006, pp 882-893*.

[14] S. Mahamud (2002). Discriminative distance measures for object detection. *Doctoral dissertation, CMU. www-2.cs.cmu.edu/ mahamud/recognition/recognition.html*.

[15] T. Moffett [Muffet] (1658). The Theatre of Insects; or, Lesser living Creatures, as Bees, Flies, Caterpillars, Spiders, Worms, etc., a most Elaborate Work in vol. 2 of Edward Topsell, The History of Four-footed Beasts and Serpents: Describing at Large Their True and Lively Figure, their several Names, Conditions, Kinds, Virtues (both Natural and Medicinal) Countries of their Breed, their Love and Hatred to Mankind. Collected out of the writings of Conradus Gesner and other authors, by Edward Topsell. Whereunto is now added, The Theater of Insects…. by T. Muff et, 2 vols. *London: printed by E. Cotes, for G. Sawbridge, T. Williams, and T. Johnson, 1658*.

[16] G. P. Nguyen, M. Worring, and A. W. M. Smeulders. Interactive search by direct manipulation of dissimilarity space. *IEEE Transactions on Multimedia. VOL. 9, NO. 7, Nov 2007*.

[17] C. B. Richard Ng, G. Lu, and D. Zhang. Performance Study of Gabor Filters and Rotation Invariant Gabor Filters. *MMM 2005: 158-162*.

[18] A. Seba (1734). Locupletissimi rerum naturalium thesauri accurata descriptio Naaukeurige beschryving van het schatryke kabinet der voornaamste seldzaamheden der natuur. *Amsterdam, 1734-1765. 4 vols. 2º. - 394 B 26-29, vol. 3, plate XXXV*.

[19] W. Smith. (1853) British Diatomaceae. *Volume 1. John Van Voorst, London: xxxiv+89pp*.

[20] S. Squire. (1998) Learning a Similarity-Based Distance Measure for Image Database Organization from Human Partitionings of an Image Set. *IEEE Workshop on Applications of Computer Vision (WACV'98), pp.88-93, 1998*.

[21] Varde, A., Rundensteiner, E., Javidi, G., Sheybani, E. and Liang J. (2007). Learning the Relative Importance of Features in Image Data. In *Proceedings of IEEE ICDE's DBRank-07*, Istanbul, Turkey, April 2007, pp. 237 - 244

[22] R. C. Veltkamp, and L. J. Latecki. Properties and Performance of Shape Similarity Measures. *In Proceedings of IFCS 2006 Conference: Data Science and Classification. July, 2006*.

[23] X. Wang, L. Ye, E. Keogh, and C. Shelton (2008). www.cs.ucr.edu/~xwang/historical_shapes/index.html (While this work is still under review, this page will be password protected. *Username* **ucr**, *Password* **12345**)

[24] *www.entomo.pl/lepidoptera/galeria_motyli/motyle_nocne.htm Visited on 26-Oct-07*.

[25] *www.temple-of-flora.com/natural_history.htm Visited on 28-Oct-07*.

[26] *www.texasarrowheads.com/ Visited on 15-Jan-08*.

Appendix A:

To reduce visual clutter and enhance the flow of the text, we avoided naming all the insects in the experiments shown in Figure 14. For completeness we do that here.

In the top section we have 6 pairs of insects, which are, from top to bottom, left to right: {*Anaea thebais*}, {*Attacus atlas*}, {*Morpho menelaus*}, {*Eurytides marcellus*}, {*Geometra papilionaria*}, {*Graphium androcles*}.

In the center section we have 6 pairs of insects, which are, from top to bottom, left to right: {*Actias maenas, Argema mittrei*}, {*Danaus plexippus, Speyeria idalia*}, {*Hemaris thysbe, Macroglossia fuciformis*}, {*Hestia lyncea, Papilio veiovis*}, {*Papilio cresphontes, Papilio machaon*},{*Troides amphrysus, Troides oblongomaculatus*}.

In the bottom section we have 2 pairs of insects, which are, from, left to right: {*Tacua speciosa, Salvanza imperialis*}, {*Urbanus proteus, Telesilaus telesilaus*}.