

High Distortion and Non-Structural Image Matching via Feature Co-occurrence

Xi Chen Tat-Jen Cham
School of Computer Engineering
Nanyang Technological University
Singapore

chenxi@pmail.ntu.edu.sg astjcham@ntu.edu.sg

Abstract

We propose a novel approach for determining if a pair of images match each other under the effect of a high-distortion transformation or non-structural relation. The co-occurrence statistics between features across a pair of images are learned from a training set comprising matched and mismatched image pairs – these are expressed in the form of a cross-feature ratio table. The proposed method does not require feature-to-feature correspondences, but instead identifies and exploits feature co-occurrences that are able to provide discriminative result from the transformation. The method not only allows for the matching of test image pairs that have substantially different visual content as compared to those present in the training set, but also caters for transformations and relations that do not preserve image structure.

1. Introduction

The task of *image matching*, as considered in the most generic context, involves determining whether a pair of images exhibit a special contextual relationship that distinguishes it from other random pairings of images. For example, human IQ quizzes regularly require the matching pairs of visual patterns that have undergone a particular spatial transformation, that is implicitly captured in other analogous pairs of *dissimilar* patterns provided as prior examples. In another example, children books and television programs often contain problems in which pairwise matching of visual images have to be performed based on more universally known contextual information, e.g. matching of two images that are of different parts of the same object.

In extending this generalized concept of image matching to machine vision, we want to highlight two key problems that are not considered in current methods:

- *Large differences of visual content between training data and actual/test data.* In image matching problems

that may be cast into an object recognition framework, the goal is to recognize an object in the image, based on training data containing objects of a similar class. This presumes that the visual appearance or features of the object in the test image is adequately spanned by training data. This is not true of cases in figure 1.

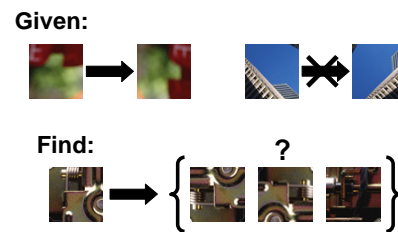


Figure 1. Learning with large differences of visual content between training data and actual/test data. Existing object recognition methods depend on training data that are visually or feature-wise similar to test data.

- *Large differences of visual content between two (correctly) matched images.* Two images may have visual content that are entirely different from each other but nevertheless semantically related, possibly in a directional manner. For example in the bottom row of figure 2, there is an obvious relationship between the upper and bottom halves of a face, but the image structure is entirely different. Similarly, for the top row in figure 2, images are systematically related in the sense that the right image is the end product of filtering the left image through a fixed sequence of complex filters. However, in all instances, it is not only difficult to find pointwise correspondences, they do not even exist.

Most of existing methods do not address these problems (they were never intended to in the first place). For example, popular classification-based recognition methods depend on extensive training data. Nevertheless, there are numerous image matching problems that involve matching two images for which the visual content of the images do not belong to some prior known class. Such problems include

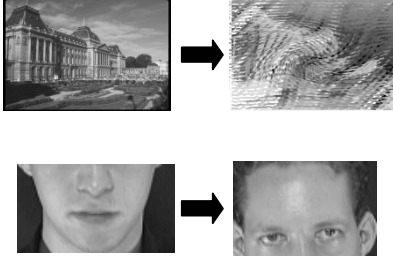


Figure 2. Large differences of visual content between “matched” images. Top row: this illustrates the result of the left original image undergoing a series of filters, resulting in the right image. Bottom row: Bottom half of a face is related to the top half. Current methods do not adequately deal with matching images with these forms of relation, particularly for test data which are also significantly biased as compared to the training data.

image mosaicing, stereo correspondence and tracking with online acquired models. These problems do not readily admit the use of classification approaches. As such, matching of image pairs have traditionally been carried out using simple metrics such as sum-of-squared differences (SSD) or histogram matching, with research focused on robustly estimating warp parameters that align image pairs.

A possible approach is to directly engineer solutions for each domain separately. For example, to solve the matching problem in figure 1, one may attempt to use affine-invariant region detectors. The problem with such an approach is that explicit transformation models have to be assumed; furthermore, constraints on the range of transformations that are implicitly defined by the training data may be lost through this approach. More challenging would be the problem in figure 2, where no existing methods are able to address.

In this paper, we make the following contributions:

- we articulate the problem of matching across severe, correspondence-breaking transformations as well as relations that do not preserve structure;
- provide a mechanism for matching test images which are visually very different from training data; and
- propose a unified framework for learning such transformations and relations directly from training data, avoiding the need to engineer domain-specific models.

1.1. Related Work

In the past, standard measures for evaluating correspondences are the sum-of-squared pixel differences (SSD), standard metrics in feature space (e.g. Euclidean is used in the SIFT framework [10]), the chamfer distance [2], the shuffle distance [18], the cross-affinity family of distances [3] and histogram measures such as χ^2 distance, Bhattacharyya distance (as used in the mean-shift framework [5]), Kullback-Leibler divergence and Earth Mover’s

Distance [14]. These measures, while intuitive, fail when used for comparing images under heavy-distortion transformations and non-similar relations as in figure 2.

Histograms of features has been successfully applied to image classification [6, 7], multimedia classification [11] and event recognition [20]. There are also methods that involve feature selection [1, 4]. However in most of these cases, the requirement is that training data is available which has similar visual content (and will generate similar histograms) to the test cases. This requirement also applies to recent classifier-based methods for object recognition involving SVM [8, 19].

The image analogies work in Hertzmann et al. [9] is semi-related in that feature-wise relationships are involved. However, image analogies deal with image synthesis and require point-to-point correspondences, whereas this paper deals with image matching and the framework does not require correspondences, even catering for matching image structures so different that correspondences may not exist.

2. Overview of Approach

At the heart of our proposed approach is the notion that images are composed of numerous small features as building blocks of the image. In spirit this is similar to existing approaches based on collections of sparse corner features, except that in our case, a fully dense and over-complete representation is used as discussed in section 3. Additionally, unlike these methods we do not attempt to characterize an image based on extracted features, since we are not interested in the images themselves *per se* – after all, the images in our test data is significantly different from those in the training data. Instead, our focus is on describing how features relate to other features *across* the relevant set of transformations or relations, where the set of relations is implicitly captured by training data provided.

A key assumption made in this paper is that under a particular set of transformations or relations, certain features in one image are present in *consistent ratios* with other features in the second image, where these features may be of different types between the two images. While there is no exact basis for this assumption, this is generally true for many transformations and relations. For example, consider a transformation comprising purely a 90° rotation followed by grayscale inversion. Features corresponding to 45° dark edges will be present in equal quantity with -45° light edges across the transformation. Similarly, in a relation of upper faces to lower faces, there will be four eye-corner features for two mouth-corner features (note: this example is used illustratively and we do not explicitly search for such domain-specific features). Hence each new image, even with substantially different visual content from those previously observed, is simply “disassembled” into features. The frequency of each feature type (described later) may then be

individually compared to frequencies of other feature types in the hypothesized matched image.

The main challenge of this approach is to determine *which feature-to-feature ratios are systematically the result of the underlying transformation / relation, and which are attributed to incidental variations in image content, clutter and noise*. Briefly, the process adopted in our framework is as follows:

1. Start with a positive training set comprising *pairs of matched images*.
2. For each feature type, measure the frequency of occurrence within each image.
3. Measure the frequency ratios between *all pairs* of feature types across each two matched images.
4. Quantize the frequency ratios and construct a normalized histogram of such ratios across all images in the positive training set.
5. Repeat with a negative training set comprising *pairs of mis-matched images*.
6. The differences between the histograms of the positive and negative test sets indicate the frequency ratios that are important in determining whether a test pair of images are matched.

The subsequent sections describe this process in more precise terms. Additionally, we also use the term “transformation” to describe both normal transformations as well as contextual pairwise relations such as two halves of a face.

3. Image Representation

In the framework of this paper, a raw pixel image I is converted into a representational form of a feature histogram:

$$I \mapsto \mathbf{h} = H(F(I)) \quad (1)$$

where h is an integer vector representing the histogram, $H(\cdot)$ is a histogramming operator and $F(\cdot)$ is a feature extraction operator. As discussed earlier, instead of adopting the recent highly popular sparse representations (*e.g.* [6]), $F(\cdot)$ in our paper involves extracting a fully-dense and over-complete feature set through convolution with a small filter bank and returns the complete filter responses. Similarly for the image with transformation:

$$\mathcal{T}(I) \mapsto \mathbf{h}' = H(F(\mathcal{T}(I))) \quad (2)$$

where $\mathcal{T}(\cdot)$ representing the unmodeled transformation.

In this way, each image may be considered as mapped to a fixed-length integer vector. In this representation, we consider each “feature type” of the image as the bin index of this histogram vector, and each “feature value” as the integer value of the associated bin.

4. Transformation Representation by Feature Co-Occurrence

Given two images I and I' which have been converted into histograms \mathbf{h} and \mathbf{h}' respectively, the problem remains on how to compare these images. In a classical approach, one might attempt to either learn $p(\mathbf{h}, \mathbf{h}'|T)$ where T is an label indicating if images are related via the transformation $I' = \mathcal{T}(I)$, or attempt to classify $(\mathbf{h}, \mathbf{h}')$ via SVMs. In either case, a large amount of training data is usually needed, and also presupposes that the test images appear similar to those in the training data. However, if the test images are significantly different, these approaches will generally perform poorly.

Instead, we adopt a component-based approach by assuming that the features of h_1, h_2, \dots, h_N in \mathbf{h} *individually* represent abstract building blocks of images, and that new images are formed through the assembly of these features. This is analogous to the observation that words are the building blocks of documents. Indeed, word-based analysis has led to the use of measures in information-retrieval research such as tf-idf [15] that involves *term frequency* and *document frequency* in measuring the discriminative value of words for topical clustering of documents. Furthermore, text-based approaches have also been successfully applied to image classification [13, 17, 12].

However in these previous approaches, these generalized “terms” are used to represent the *content* of the images, rather than the *transformation* which is what we are after. In order to extend the representation to the representation of the transformation, we consider **pairwise ratios of cross-image features**. In other words, given

$$\mathbf{h} = [h_1, h_2, \dots, h_N] \quad \text{and} \quad \mathbf{h}' = [h'_1, h'_2, \dots, h'_N], \quad (3)$$

the transformation may be represented in the form

$$\mathcal{T} \sim R = \begin{bmatrix} h_1:h'_1 & \dots & h_N:h'_1 \\ h_1:h'_2 & \dots & h_N:h'_2 \\ \dots & \dots & \dots \\ h_1:h'_N & \dots & h_N:h'_N \end{bmatrix} \quad (4)$$

Here we can consider R as a $N \times N$ table, which we call the *Cross-feature Ratio Table* (CRT). The r_{ij} entries of the CRT effectively captures the co-occurrence statistics of such pairs of features. An alternative interpretation is that these pairs of features are meta-terms, and the statistics capture the meta-term frequency within a pair of matched images. See figure 3.

5. Transformation Comparison

The goal of this paper is that given an ordered pair of images, determine if it belongs to:

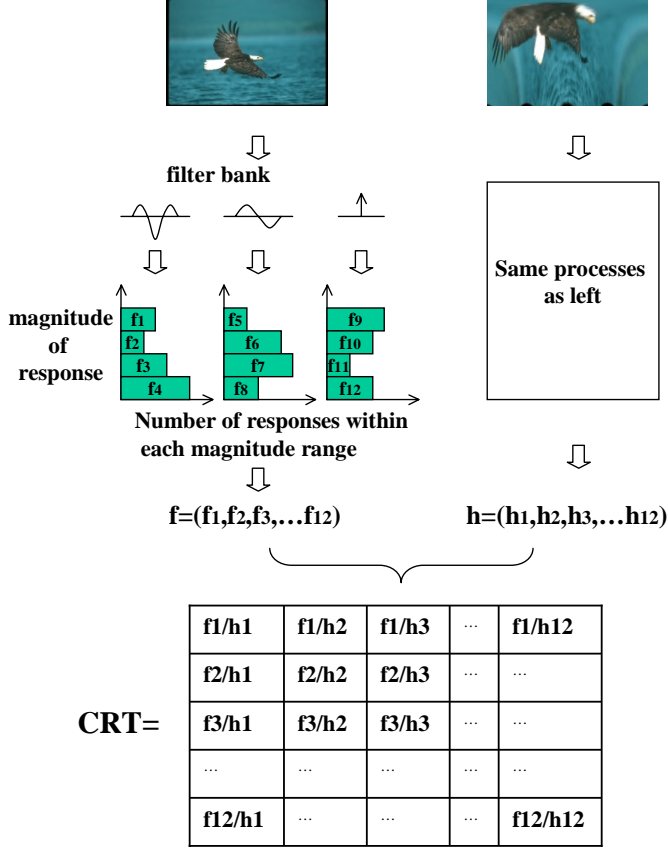


Figure 3. Overview of process for creating the Cross-feature Ratio Table for a single pair of images.

- *Positive transformation class*, which captures the range of transformations that we want to match under; or
- *Negative transformation class*, which could either represent:
 - non-transformations (*i.e.* the original image and the source of the transformed image are different), or
 - a range of transformations that we want to consider mismatch. For example, we may want to distinguish between image rotations of 45° to 90° versus -45° to -90° .

6. Discriminative Analysis of Cross-Feature Ratios

The approach of this paper to transformation comparison is to effectively determine which of the features the cross-feature ratios r_{ij} are useful in discriminating between the positive and negative transformation classes.

Suppose we are given positive and negative *image pairs* that form the training dataset. By computing the separate

CRT's for these image pairs, we can compute the separate distributions for each of the CRT entries r_{ij} converted to fractional form:

$$H_{ij}^+(k) = p(\alpha(k) < r_{ij} < \alpha(k+1)|T) \quad (5)$$

$$H_{ij}^-(k) = p(\alpha(k) < r_{ij} < \alpha(k+1)|-T) \quad (6)$$

where $H_{ij}^+(k)$ and $H_{ij}^-(k)$ are normalized histogram (probability distribution) functions for the positive and negative set respectively, k is the index to a set of quantized ratios. Here, we ignore cross-feature response ratios of $0 : 0$ (*i.e.* no instances of either type of features exist in the pair of images) as these do not contribute in any fashion to discrimination.

The ideal scenario will be that some of the H_{ij}^+ and H_{ij}^- provide strong discrimination between cross-feature ratios from the positive and negative transformation classes. In order to evaluate this discrimination, we measure the χ^2 divergence to estimate the relative discriminative power of the cross-feature ratios:

$$\begin{aligned} w_{ij} &= \chi^2(H_{ij}^+, H_{ij}^-) \\ &= \sum_k (H_{ij}^+(k) - H_{ij}^-(k))^2 \\ &, \text{ for } 1 \leq i, j \leq n \end{aligned} \quad (7)$$

The w_{ij} factors will be used to weight the cross-feature response ratios when computing similarity in the subsequent section.

7. Transformation Matching via Weighted Cross-Feature Ratios

Given a pair of test images, we can compute the associated CRT R . For simplicity, R is expressed in the form of H_{ij}^r , although the individual histograms are simply null or delta functions because for only a single pair of images, there is at most one entry for r_{ij} , and hence at most one non-zero bin contain 1.

The combined distance between the test pair CRT distribution H_{ij}^r and the positive transformation class CRT distribution H_{ij}^+ is defined as:

$$D^+ = \sum_{i,j} w_{ij} (1 - (H_{ij}^r \bullet H_{ij}^+)) \quad (8)$$

where “ \bullet ” refers to the inner product for which the distributions H_{ij}^r and H_{ij}^+ are treated as vectors. Similarly, we can define a distance to the negative transformation class

$$D^- = \sum_{i,j} w_{ij} (1 - (H_{ij}^r \bullet H_{ij}^-)) \quad (9)$$

The distances D^+ and D^- are used in deciding if the test image pair belongs to the positive or negative transformation class. The decision rule is given by

$$c = \text{sgn}(D^- - \lambda D^+) \quad (10)$$

where $\text{sgn}(x)$ is the sign function given by

$$\text{sgn}(x) = \begin{cases} 1, & \text{for } x > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

λ is a constant parameter that is learnt from training. When $c = 1$, the image pair is computed as a match, while if $c = -1$ then the image pair is deemed mismatched.

8. Experiments

We tested our approach on Corel Stock Photo Library2 dataset, using 48 features derived from 6 simple filters. The transformations include random projective transforms in a parameter space, and Adobe Photoshop special effects with large spatial and intensity displacements. Additionally, we test relation-based matching using the ORL face dataset by dividing the face images into separate left/right and top/bottom halves.

Note that the framework for training and classifying the matches are exactly the same across all the experiments. This is one of the key goals of our work – to provide a *unified framework for learning such relationships directly from training data, without the need to create domain-specific models* that require manual intervention and possibly extensive engineering.

8.1. Dataset Construction

Corel Stock Photo Library2 includes 186 categories of photos, such as African Wildlife, Autumn in Maine and Freestyle Skiing, etc. Each category has 100 color images of size 768×512 or 512×768 , with quite different visual contents.

We randomly selected 20 categories as source for the training data, and another 20 categories for testing data. All the images are resized to 25% of original size by bilinear interpolation, and changed to grayscale images.

We experimented with two differently constructed datasets:

- **Type (1) dataset.** The positive set is composed of image pairs related by transformation, and the negative set is composed of unrelated image pairs, possibly containing different visual content.
- **Type (2) dataset.** The positive set is composed of image pairs related by a subset of transformations, while negative set is composed by image pairs related by a different subset of transformations (*i.e.* the transformations in the positive set have a different range of parameters compared to those in the negative set).

8.2. Filter Bank

The raw filtered features are derived from 6 simple filters: Delta, LoG and 4 oriented edge filters with 45° steps, each filter size is 3 by 3. The filter responses are divided into 8 ranges evenly distributed between -256 and 256, corresponding to 8 different features for each filter. One feature response is the number of responses of one filter fall in one range within the whole image. So totally there are 48 features used.

8.3. Results for Geometric Transformation

The geometric transformation is composed of a random but moderately small projective transform component followed by a random large rotation component. The projective component is defined by 4 pairs of reference and corresponding points. The four reference points are located at (50,50),(50,-50),(-50,50) and (-50,-50), and each corresponding point is randomly located within a 20 by 20 box centered at the reference point. The rotation component is a pure rotation of a random degree centered at (0,0). All the coordinates are using pixel as unit. For doing image transformation, the origin is at the center of that image.

For the type (1) dataset, the transformations comprise a small random projective component followed by a random rotation between 0° and 180° . For either the training set and testing set, we randomly chose 10 images from each of the 20 categories, cropped the 96×96 center part as reference images, resulting in 200 reference images. We applied 5 random geometric transformations on each reference image to obtain 5 corresponding images. Hence there are a total of 1000 matched image pairs in each (training and testing) positive data set. Furthermore, we created 2000 mismatched image pairs for each negative dataset, where each reference image has 10 transformed but mismatched corresponding images. Some positive image pairs in training set are shown in figure 4. The corresponding ROC curves is shown in figure 5.

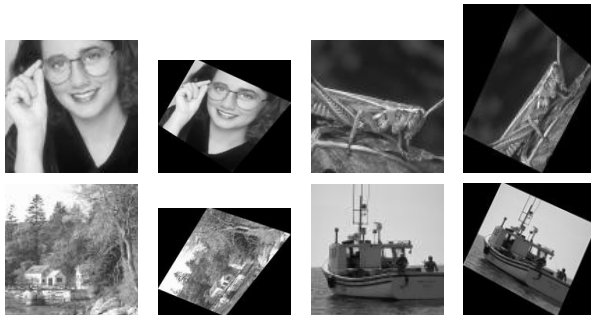


Figure 4. Some of the positive image pairs in training set are shown.

We also tested our approach on a type (2) dataset where the reference images are the same as type (1) dataset. For

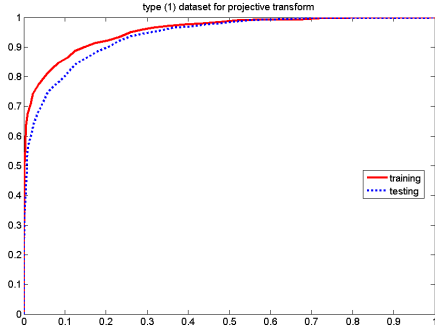


Figure 5. ROC curves for the experiment on a type (1) dataset involving geometric transformations

positive image pairs, the transformations comprise a small random projective component followed by a random rotation between 60° and 90° , while for negative pairs, the rotation is set randomly between 120° and 150° . For each reference image, we apply 5 random transformations, resulting in 1000 image pairs for both positive and negative data sets. The ROC curve for type (2) dataset is shown in figure 6. The resultant χ^2 discriminative weights across CRT entries are shown in figure 7.

The classifications results were good, which is somewhat surprising since the proposed framework did not in any way directly model the spatial transformations. More significantly, the strong results for the type 2 dataset would not be possible just by naively applying affine-invariant region detectors, as *both* positive and negative image pairs are related by projective transformations – the difference lies in the range of transformations that are only implicitly captured in the training data.

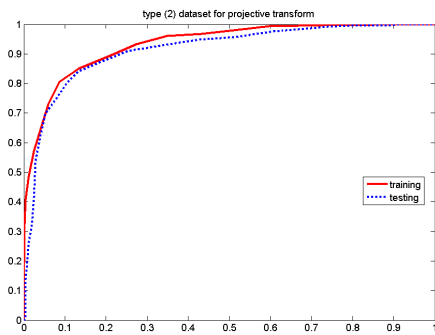
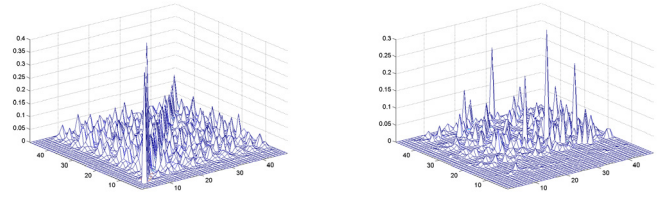


Figure 6. ROC curves for the experiment on a type (2) dataset involving geometric transformations

8.4. Results for Adobe Photoshop Special Effects

We made use of two Adobe Photoshop effects to provide the image transformations in this experiment. The first effect is graylevel histogram equalization followed by the



type(1) dataset

type(2) dataset

Figure 7. χ^2 discrimination across CRT entries. Each bin in the 2D plane corresponds to a bin in the CRT table, while the peaks denote the importance of the corresponding ratio in a discrimination task.

Wave distortion, the second is intensity inversion followed by the *Rough Pastels* and the *Twirl* distortions. For both training and testing set, 50 images were randomly chosen from each of 20 categories, resulting in 1000 reference images in each set.

For both effects, we constructed a type (1) dataset, *i.e.* we applied the effect to all reference images through a batch process, producing 1000 positive image pairs. We also created 2000 negative image pairs by selecting 2 mismatched transformed images from each reference image.

Three reference images and their corresponding transformed images by effects 1 and 2 are shown in figure 8. The ROC curves for the type (1) dataset for effects 1 and 2 are shown in figures 9 and 10 respectively.

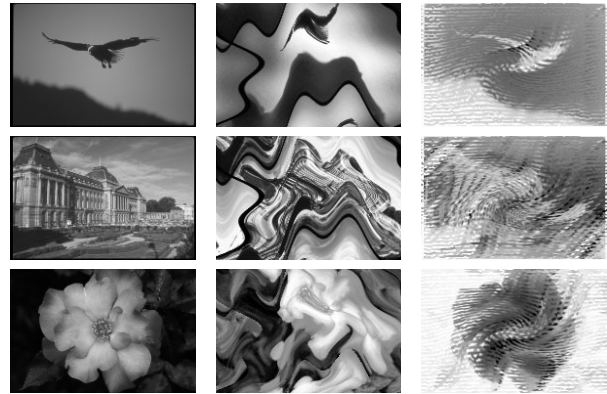


Figure 8. Left column: reference images in training set; middle column: corresponding transformed images by effect 1; right column: corresponding transformed images by effect 2

Two other image matching methods, SSD and classical χ^2 distance of graylevel histogram, are also tested on the dataset for comparison. Only the results on training data are shown, because no training is required, and the results are very similar for both training and testing datasets.

We also constructed a type (2) dataset for the two Adobe Photoshop effects. The reference images are the same as those used in type (1) dataset, but this time the positive dataset includes 1000 matched image pairs by effect1, and the negative dataset includes 1000 matched images pairs by

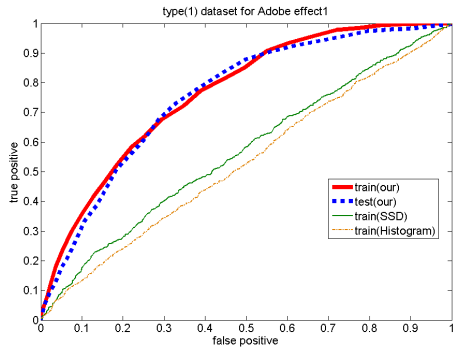


Figure 9. ROC curves for Adobe Photoshop effect1 transformation of type (1) dataset.

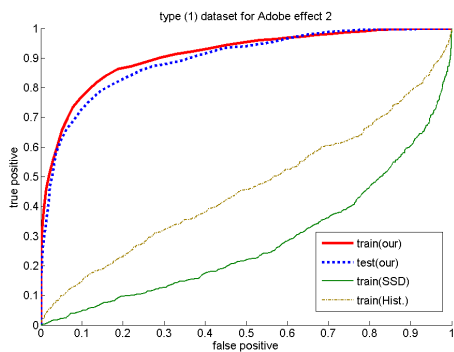


Figure 10. ROC curves for Adobe Photoshop effect2 transformation of type (1) dataset.

effect2, the reference images are the same. The ROC curves for type(2) dataset is shown in figure 11.

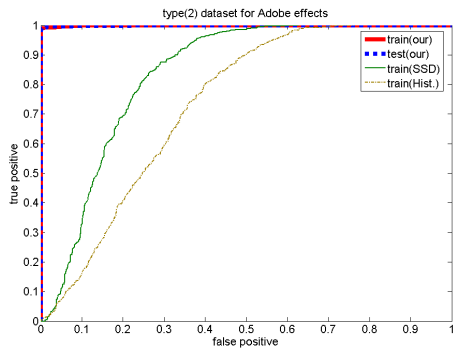


Figure 11. ROC curves for Adobe Photoshop effect2 transformation of type (2) dataset.

Despite the variation in the quality of the results, they are still surprisingly impressive, given the severity of the transformations. Note that the images themselves are even difficult for humans to distinguish, and correspondences are hard to recover, if at all possible.

8.5. Results for Partial Face Matching

We made use of the ORL face dataset [16] as a basis to create multiple training and test sets for our partial-face matching experiments.

In the first experiment, we attempt to validate the method by testing if it was able to determine image pairs comprising complementary parts of face (i.e. parts that together form a complete face). A positive training set is created by taking images from half the individuals in the ORL dataset, and dividing the images into top and bottom halves as matched pairs. To compound the problem, half of these pairs are randomly selected and their relations reversed (i.e. half the pairs have top→bottom relation, while the other pairs have bottom→top relation). See figure 12. A negative training

Positive:



Negative:



Figure 12. Dataset for Partial Face Matching Experiment 1

set is generated where for each image pair, one image is half of the face, whereas the second image is either a copy of the first image, or a random selected image from the Corel database. The test sets are similarly constructed, *except that the individuals selected in the test sets do not appear in the training sets*. Note that the ORL dataset is not specially aligned, and therefore images in the top and bottom halves do not have aligned features, nor are all the features consistently present in the same half (e.g. tip of nose may appear in either half). See figure 13 for results.

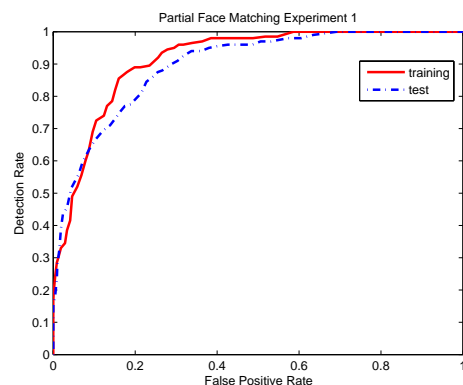


Figure 13. ROC Curve for Partial Face Matching Experiment 1

While equivalent results may be achieved if half-face models are used in matching, the results indicate that our

proposed generic framework is able to deal with the matching without domain-specific models.

In the second experiment, we attempt to evaluate if the method was capable of distinguishing if pairs of images comprising (a) left and right halves of a face, *belong to the same individual*, and similarly for (b) top and bottom halves of a face. More significantly, the testing phase involves images of *individuals who were not used in the training phase*. As before, a positive training set is created by taking images from half the individuals in the ORL dataset, and dividing the images into two halves as matched pairs (experiment 2(a) involves left/right halves, while 2(b) involves top/bottom halves). The negative training set involved randomly swapping equivalent right (or bottom) halves of images such that pairs of images are not extracted from images of the same individual. The results are shown in figure 14.

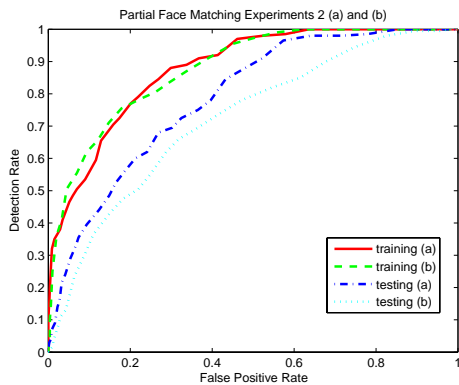


Figure 14. ROC Curves for Partial Face Matching Experiment 2

The results of the final experiment are somewhat surprisingly good, even though the actual detection rates are only moderate. The experiment demonstrated that matching of lower faces and upper faces belonging to the same individuals can be carried out, *even though the individuals have not been encountered before in the training data*. Moreover, the results were obtained using a generic framework, without the need for domain-specific face models and classifiers.

9. Conclusions

A novel approach for determining if a pair of images match each other under the effect of heavy-distortion transformation or non-structural relation. The co-occurrence statistics between features across a pair of images are learned from a training set comprising matched and mismatched image pairs – these are expressed in the form of a cross-feature ratio table. The proposed method does not require feature-to-feature correspondences, but instead identifies and exploits feature co-occurrences that are able to provide discriminative result from the transformation. The method not only allows for the matching of test image pairs

that have substantially different visual content as compared to those present in the training set, but also caters for transformations that do not preserve image structure. Experimental results indicate that this matching process not only clearly outperforms the classical methods of SSD and histogram matching, but are able to match images under extreme transformations and non-structural relations.

Acknowledgements

This work was carried out in and supported by the Centre for Multimedia and Network Technology (CeMNet), NTU.

References

- [1] S. Avidan. Joint feature-basis subset selection. In *CVPR*, 2004.
- [2] H. Barrow, J. Tenenbaum, R. Bolles, and H. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *IJCAI*, pages 659–663, 1977.
- [3] X. Chen and T. Cham. Learning feature distance measures for image correspondences. In *CVPR*, volume 2, pages 560–567, 2005.
- [4] R. Collins and Y. Liu. On-line selection of discriminative tracking features. In *ICCV*, 2003.
- [5] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, pages 142–149, 2000.
- [6] K. Grauman and T. Darrell. Efficient image matching with distributions of local invariant features. In *CVPR*, 2005.
- [7] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005.
- [8] B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machines: Global versus component-based approach. In *ICCV*, 2001.
- [9] A. Hertzmann, C. Jacobs, N. Oliver, B. Curless, and D. Salesin. Image analogies. In *SIGGRAPH*, Los Angeles, CA, 2001.
- [10] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
- [11] P. Moreno, P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *NIPS*, 2003.
- [12] D. Nistér and H. Stewénus. Scalable recognition with a vocabulary tree. In *CVPR*, New York, NY, 2006.
- [13] S. Paek, C. Sable, V. Hatzivassiloglou, A. Jaimes, B. Schiffman, S.-F. Chang, and K. McKeown. Integration of visual and text based approaches for the content labeling and classification of photographs. In *MMIR*, 1999.
- [14] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover’s distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000.
- [15] G. Salton and M. Smith. On the application of syntactic methodologies in automatic text analysis. In *SIGIR*, pages 137–150, 1989.
- [16] F. Samara and A. Harter. Parameterisation of a stochastic model for human face identification. In *WACV*, Sarasota, FL, 1994.
- [17] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, Nice, France, 2003.
- [18] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *ICCV*, 2001.
- [19] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.
- [20] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *CVPR*, 2001.