Localized Matching Using Earth Mover's Distance Towards Discovery Of Common Patterns From Small Image Samples

Hung-Khoon Tan and Chong-Wah Ngo

Department of Computer Science
City University of Hong Kong

Abstract

This paper proposes a new approach for the discovery of common patterns in a small set of images by region matching. The issues in feature robustness, matching robustness and noise artifact are addressed to delve into the potential of using regions as the basic matching unit. We novelly employ the many-to-many (M2M) matching strategy, specifically with the Earth Mover's Distance (EMD), to increase resilience towards the structural inconsistency from improper region segmentation. However, the matching pattern of M2M is dispersed and unregulated in nature, leading to the challenges of mining a common pattern while identifying the underlying transformation. To avoid analysis on unregulated matching, we propose localized matching for the collaborative mining of common patterns from multiple images. The patterns are refined iteratively using the expectation-maximization algorithm by taking advantage of the 'crowding' phenomenon in the EMD flows. Experimental results show that our approach can handle images with significant image noise and background clutter. To pinpoint the potential of Common Pattern Discovery (CPD), we further use image retrieval as an example to show the application of CPD for pattern learning in relevance feedback.

Key words: Common Pattern Discovery, Earth Mover's Distance, Localized Matching, Local Flow Maximization, Expectation-Maximization

Email addresses: hktan@cs.cityu.edu.hk, cwngo@cs.cityu.edu.hk (Hung-Khoon Tan and Chong-Wah Ngo).

1 Introduction

Huge amount of visual information in the form of digital images and video databases is generated everyday. Extracting visually common patterns from images is becoming increasingly important for various multimedia applications. The mined patterns can serve as the entry points for efficient browsing of large visual databases, while enable effective clustering and search. Common Pattern Discovery (CPD) can be regarded as a superset problem of image registration and pattern detection as shown in Figure 1. Given two images I and J, image registration finds the best transformation T that aligns I and J. Pattern detection, in addition to finding the optimal T, locates the subimage J^* of J that best matches I through the transformation T. CPD extends both problems to find the best match of a subimage I^* in I with J^* through an optimal but unknown transformation T. Compared to registration and detection, CPD has no knowledge of I^*, J^* and T which need to be simultaneously optimized, leading to a dramatically inflated search space. CPD can be extended to the multiple image case as shown in Figure 1(c). In general, more images lead to more visual evidence for the discovery of common patterns. Given a set of N images, $I_{i=1...N}$, the task of CPD is to find the subimages, $I^*_{i=1...N}$ and their transformation parameters $T_{i=1...N}$ which maximizes a particular similarity function $\mathcal{H}$. CPD can be formulated as follows

$$\{T_i, I_i^*\}_{i=1...N} = \underset{\{T_i, I_i^*\}_{i=1...N}}{argmax} \; \mathcal{H}(T_1(I_1^*), ..., T_N(I_N^*)) \tag{1}$$

CPD, by definition, is related to but different from the recent research in visual category recognition [1]. CPD is a multiple image matching technique that performs 'exact' object extraction while visual category modeling learns the visual models of object categories that may cover a wide spectrum of visual appearances. Although visual category recognition also looks for visually consistent patterns in a group of images, it is generally a *learning* process which uses a *training set* to capture the variability in appearances among the objects either through discriminative or generative modeling. On the other hand, CPD is a *matching* technique that finds deformed replicates of the same object from a *small* set of unconstrained images. It performs vigorous search to mine common patterns which might have undergone affine and photometric transformation from images with background clutter. Despite the fundamental difference, in practice CPD can be used as a pre-processing step for visual category recognition. A good model for recognition can be learnt from common patterns discovered in a training set.

Generally, CPD performs pattern mining in search of the optimum matching at the sub-image level. It searches for an unknown subset of primitives from an image that best matches all equally unknown subsets of all other images, in

Fig. 1. (a) Image registration: Given a template (first image), find the affine transformation to the second image (second image). (b) Pattern Detection: Given a template (first image), mine the sub-image and the transformation parameters from a target (second) image. (c) Common Pattern Discovery: Without any prior information of the common pattern, mine the unknown common sub-images given a set of images.

terms of the appearance and geometric consistency. A fundamental problem why the problem is challenging is that the visual data is unstructured and unordered, leading to ambiguity in grouping perceptually meaningful patterns. The problem becomes even harder when input images also contain variations in terms of viewpoint, affine transformations (translation, rotation, stretching and scale), photometric transformations (color, illumination and shading) and occlusion.

The challenges in CPD include feature robustness, matching robustness and noise artefacts. For feature robustness, matching can be performed either *sparsely* on a set of points at visually interesting locations picked by feature detectors [2], or *densely* on a set of blocks from grid partitioning or a set of regions generated from the segmentation of the image space, respectively. Having a set of robust features against the variations of object appearance and transformation is always difficult. Points, sampled at prominent locations, can only sketch the outline of a common pattern but lack sufficient details especially on homogeneous parts, thus sacrificing completeness. Blocks, uniformly and densely sampled from whole image space, are not tolerant to scale change due to rigid partitioning. Regions, adaptively and coherently segmented, preserve completeness for matching. However, robustness is a major concern because image segmentation is hardly perfect. Under such circumstances, the choice of matching strategies becomes critical in order to overcome the structural disparity due to image segmentation.

Depending on the mapping constraint being imposed, matching can be categorized into one-to-one (O2O), many-to-one (M2O), one-to-many (O2M) and

many-to-many (M2M) matching. For CPD, O2O is the de facto matching scheme [3–5,10] because unique pair-wise correspondences are useful to estimate the affine transformation between two common patterns. O2O, however, is sub-optimal particularly for region matching since the erroneously segmented regions cannot be effectively matched. For instance, due to imperfect segmentation, one region in an image ideally should match to a collection of broken regions in another image. To tolerate structural perturbation due to image segmentation, M2M appears as a more generalized matching strategy, with O2O, O2M and M2O as its special cases. M2M allows bidirectional partial matching and thus is able to adaptively correlate two sets of fragmented regions. For CPD which searches for common patterns at sub-image level, M2M still poses serious challenge since the matching patterns of M2M could be chaotic and unregulated in nature and warrants further investigations.

Noise artefacts could influence the decision of matching in several ways. Confusion may arise when several semantically unrelated regions form a set of well-aligned patterns that by chance well-correlates across the query images. This occurs frequently when the background contains significant clutter and a local optimization algorithm is used to solve the correspondence problem. In the presence of noise, matching, particularly O2O, would typically tolerate irrelevant correspondences to achieve overall consistency, or removes relevant but noise-inflicted correspondences. Noise artifacts could also exist in terms of common background. For example, the object car always co-occurs with the background concepts like road and people. In fact, by definition, the common background qualifies as a common pattern in its own right. In this case, the use of negative images as additional information is necessary to guide the mining of common patterns.

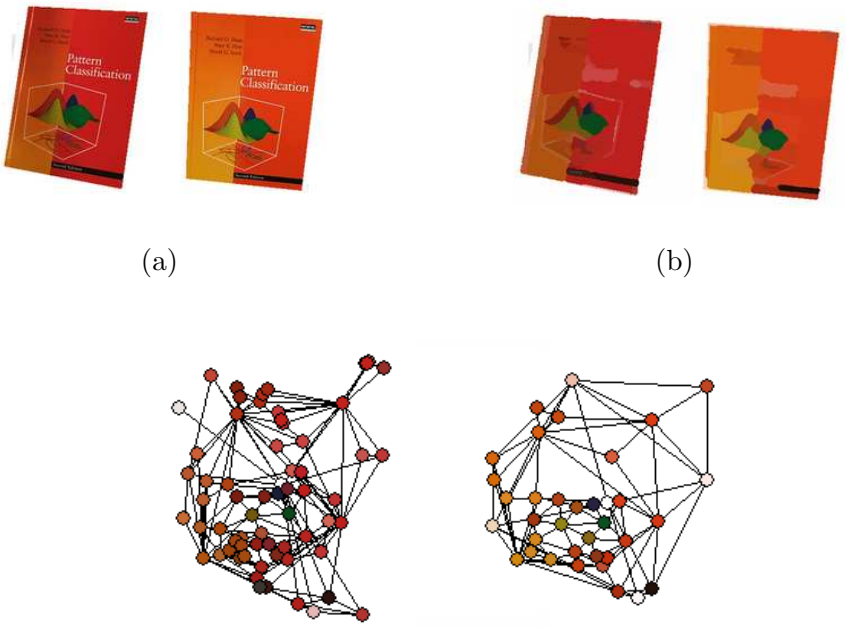
The remaining of the paper is organized as follows. Section 2 describes related works while Section 3 gives an overview of our approach. Section 4 presents our approach on incorporating the visual and spatial information for similarity measure based upon Earth Mover's Distance (EMD) while Section 5 further describes the employment of local EMD flows for mining common patterns. The proposed approach, namely Local Flow Maximization (LFM), iteratively mines the position and scale of common pattern across multiple images through EM algorithm. Section 6 presents our experimental results, while section 7 discusses the application of CPD to image retrieval. Finally, Section 8 concludes this paper.

2 Previous Works

Previous works on CPD can roughly be categorized into three major directions. The first direction mainly focuses on graph-based techniques. Images are

first segmented into constellations of homogeneous regions and then converted into graph representations such as the Attributed Relational Graph (ARG) as shown in Figure 2. CPD can then be solved as a subgraph isomorphism problem. In [3,4], Hong and Huang use a linear combination of graph model components to handle the variations in the common patterns. Expectation-maximization (EM) is then used to iteratively find the model parameters. In [5], Jiang and Ngo proposed a backtrack depth first search algorithm to mine for the maximal common subgraph from a set of ARGs. In [6], multiscale segmentation tree is used as the representation of choice where geometric and photometric attributes are taken into account when looking for the maximum common sub-trees. Although the graph-based approaches provide a ready and intuitive framework for CPD, its effectiveness is undermined by the ambiguities resulting from segmentation. During this process, homogeneous regions could be over-segmented into smaller pieces, and non-homogeneous regions could be erroneously merged into a single region, as a result of imaging variations such as shading, scale, viewpoint and illuminations, or heuristic segmentation settings. Figure 2 shows the structural inconsistency of the ARGs generated by the segmentation step owing to illumination variation. In our previous work [7], we show that EMD matching on the ARG of a segmented image is less sensitive to the structural inconsistencies arising from image segmentation. The common pattern is iteratively discovered as the local region where the flows of the EMD maximize using EM.

The second direction adopts point-to-point (P2P) matching and avoids image segmentation. Feature points which are similar in appearance and geometrically consistent are extracted as the correspondences across images. On this front, the work by Berg, Berg and Malik [8] provides the state of the art P2P matching between a pair of images. They pose the correspondence between two sets of points as an integer quadratic programming problem, where the cost function is based on the local appearance and geometric distortion between pairs of corresponding points. The matching technique has been shown to be successfully adapted for pattern detection (with exemplars manually selected). However, when extended to automatic model building, a task equivalent to CPD, it requires a large number of exemplars to statistically determine significant points. Another important work is on semi-local affine parts for object recognition [9] where groups of geometrically semi-local affine regions are mined. Triplets of local affine regions (ellipse) form the basic matching units. Triplets having similar appearance and consistent geometry in terms of the ellipse orientations are matched and extracted. The matching pairs are then locally grown by looking for similarly consistent neighbors. A validation set is used to rank the local structure according to its repeatability. However, there is an implicit assumption that the background is non-repeatable. In summary, it is unknown how robust these approaches are when the number of training images are as few as 3 to 5 images, especially with the presence of background clutter. In [10], Jiang and Ngo attempt CPD using a small set of training



(a) (b)

(c)

Fig. 2. (a) Two original images with slight variation in illumination. (b) Segmented version of (a) consists of several image segments in different colors. (c) Attributed Relational Graphs (ARG) of (b). Each node of the ARG represents a segment of the image, and the color of that segment is the attribute of the node. There is an edge connecting two nodes if and only if the corresponding two segments are adjacent to each other in the image. Slight variation in illumination results in inconsistent ARG structures and connectivity. These inconsistencies would be further aggravated when affine transformations are present.

images. Images are partitioned into grids of blocks and a color histogram is extracted for each block. Given two images, a bipartite graph is constructed with two sets of blocks. The Maximum Weighted Bipartite Graph (MWBG) matching algorithm is employed for finding the block correspondences, while procrustes analysis is adopted for finding the optimum transformation. The optimizations of matching and transformation are carried out iteratively until convergence.

Both the graph-based and point-based methods advocate matching as its underlying strategy and agree that the common pattern is constructed from multiple visual parts. The third direction, known as multiple instance learning (MIL) adopts a different approach and assumes that the common pattern can be succinctly represented by a single feature vector. Sample features are pooled from the image set to track the best point in the feature space to represent the common pattern. In most MIL algorithms [11–16], the training images are labeled as positive and negative bags, respectively, depending on

the existence of the common pattern. The training images are partitioned into segments, and low-level features are extracted to form the bag-of-features. The common pattern is found by locating a feature point near to most positive bags but far from the negative bags. One popular technique is the Diverse Density (DD) [11] where gradient ascent is employed to locate the optimal feature point. MIL requires a large amount of training images for reliable statistical analysis. In order to mine patterns that is invariant to various transformations, a large and diverse set of features is extracted but this inevitably results in the increase of noise in the feature pool. The estimation of an optimal feature point becomes highly difficult in this setting. In addition, the capacity of a feature vector to highlight the multiple variations in the common pattern is questionable. MIL only operates in the feature space, and therefore suffers from an over-reliance on the features for the description of a common pattern. Compared to MIL, matching approaches are relatively robust for the capability of excluding noises by considering both feature and geometric consistencies as shown in our experiment later.

Other recent approaches are [17] and [18]. In [17], data mining approach has been employed to rapidly discover frequent spatial keypoint configurations from tens of thousands of candidates. In the approach, keypoints are initially soft-quantized into discrete visual keywords and then frequent pattern mining, specifically the *Apriori* algorithm, is employed to discover groups of keypoints that are found to always co-exist within a localized neighborhood. In [18], a random partitioning scheme is adopted where the image spaces are randomly partitioned over many rounds. The subimages are then matched and the series of popular images are aggregated to produce overlapping blocks known as 'voting map'. The votes accumulated at each block are influenced by its size as well as the number of correspondences to other blocks. The areas with high concentration of highly popular blocks thus constitute the common pattern.

In this paper, we extend on our previous work in [7]. We employ negative samples (cf. Section 4.2) as a tool to overcome the common background problem where background patterns that always coexist with the pattern of interest are indistinguishable in the absence of any supplementary information. In [7], the initial positions of the common patterns are estimated using the mean of the flows of EMD matching conducted across the images. In this paper, we use a weighted Parzen-window (cf. Section 5.3) to take into account the density of the flows to derive a better initialization. In addition, a speedup technique has been proposed (cf. Section 5.2) by tracking the candidate common patterns to the best possible location without re-estimating the scale during each iteration. Finally, more comprehensive experiments are carried out to further confirm the effectiveness of the approach. This includes the application of CPD to image retrieval tasks to illustrate its potential for vision-based tasks (cf. Section 7).

3 Overview of Proposed CPD

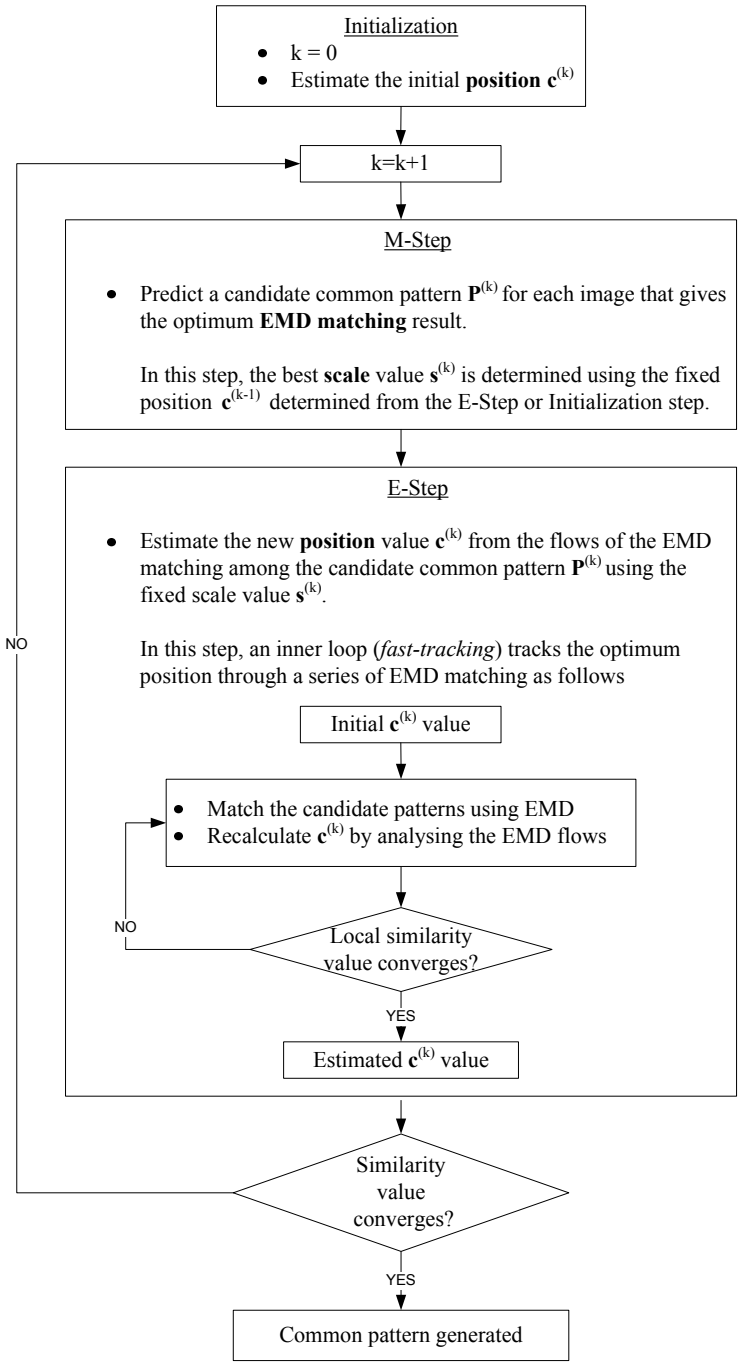


Fig. 3. Our proposed method. The scale and centroid of the candidate common pattern are updated iteratively during the maximization and expectation stage, respectively.

In this paper, we propose CPD with image-segmented regions as units and EMD as the M2M matching strategy. Region is chosen for its greater potential in describing a complete common pattern. The corresponding robustness

issue of region units is tackled with EMD which simulates M2M matching to map erroneously segmented regions, resulting in a slightly more involved matching pattern. The problem is formulated as a missing data problem, and solved under expectation-maximization (EM) framework with the regions as the observations, and the scale and centroid of a common pattern as missing data, through the analysis of the flows of regions among images.

An overview of the proposed framework is shown in Figure 3. First, the initialization step predicts the initial position values $\mathbf{c}^{(0)}$ of the common patterns (cf. Section 5.3). Then, the best scales $\mathbf{s}^{(k)}$ and positions $\mathbf{c}^{(k)}$ of the candidate common patterns $\mathbf{P}^{(k)}$ are updated iteratively during the maximization (cf. Section 5.1) and the expectation (cf. Section 5.2) steps, respectively. The estimation of $\mathbf{c}^{(k)}$ is essentially a tracking process based upon a sucessive chain of EMD matchings. In our approach, matching is performed locally instead of globally, where candidate common patterns $\mathbf{P}^{(k)}$ are first *predicted* from each positive example and then matched as a whole across multiple images. Localized matching is more tolerant to noise since the correspondences from multiple regions from the predicted patterns are considered jointly, while avoiding irrelevant correspondences from irrelevant regions which have been ruled out by the prediction. At such granularity, the goodness of matching is determined collectively as a bag of correspondences rather than assessing each match pair individually. The predicted common patterns are iteratively refined in subsequent iterations until an optimum solution is found.

To regulate the chaotic correspondence fabric, negative cues and spatial information is loosely embedded into the similarity measure for matching, thus adding more descriptive power to the underlying features. This encourages more flow transfer among the common patterns and less interactions with irrelevant nodes. Tracking the detailed transformations of the common patterns in different images aids the pattern mining process. However, in this paper we consider only the scale and translation on common pattern during mining. Knowing fine transformation parameters such as rotation, stretching and reflection is not necessary since the ultimate goal is to locate the common pattern. Obviously, the biggest challenge in localized matching is to effectively pick the best candidate pattern for matching.

4 Many-to-Many Matching with Earth Mover's Distance

4.1 Earth Mover's Distance

EMD measures the distance between two weighted point sets as a transportation problem [19]. A point set is normally referred to as a signature. EMD

strives to find the minimum amount of "work" to transport the weights from the source (supplier) to the destination (consumer) signature. In the transportation problem, a group of suppliers is required to provide a given amount of goods to a group of customers, each with a given limited of capacity to accept goods. For each supplier-customer pair, the cost of transporting a single unit of goods is given. The transportation problem is to find the least-expensive flow of goods from the suppliers to the consumers that satisfies the consumers' demand. EMD has been successfully adopted for various applications including image retrieval [19], database navigation [20], and low-level image processing [21].

EMD exhibits M2M matching for its capability in transferring partial weights between any two signatures. In our context, each signature refers to a set of segmented regions. Each region is represented by (p_i, w_i), where p_i is the mean color value and w_j is the fraction of pixels in region i. From the perspective of M2M, EMD can distribute and transfer the weight of a region to multiple broken regions in the destination signature. Conversely, a region can accept weights from multiple regions of the source signature. Give two signatures $\mathcal{S} = \{(s_1, w_{s_1})...(s_m, w_{s_m})\}$ and $\mathcal{D} = \{(d_1, w_{d_1})...(d_n, w_{d_n})\}$, EMD is formulated as:

$$\mathrm{EMD}(\mathcal{S}, \mathcal{D}) = \min \mathrm{WORK}(\mathcal{S}, \mathcal{D}, \mathbf{F}); \tag{2}$$

$$\mathrm{WORK}(\mathcal{S}, \mathcal{D}, \mathbf{F}) = \sum_{i=1}^{m}\sum_{j=1}^{n} dist(i,j) \times flow(i,j) \tag{3}$$

subject to the following constraints:

$$flow(i,j) \geq 0, \qquad 1 \leq i \leq m, 1 \leq j \leq n$$

$$\sum_{j=1}^{n} flow(i,j) \leq w_{s_i}, \qquad 1 \leq i \leq m$$

$$\sum_{i=1}^{m} flow(i,j) \leq w_{d_i}, \qquad 1 \leq i \leq n$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} flow(i,j) = \min(\sum_{i=1}^{m} w_{s_i}, \sum_{j=1}^{n} w_{d_i})$$

where $\mathbf{F}$ is a bag of flows with $flow(i,j)$ representing the flow from region s_i in $\mathcal{S}$ to region d_j in $\mathcal{D}$ and $dist(i,j)$ is the dissimilarity measure between s_i and d_j.

There are several properties of EMD that is favorable for CPD. First, it supports partial matching. Two signatures with different numbers of regions, each having different size, can be aligned under many-to-many matching with EMD. By one-to-one matching, the mapping between regions is incomplete and can be unpredictable. In contrast, EMD enlightens the chance of matching a properly segmented region to a collection of broken regions in another image. Sec-

ond, it can be proven that if the ground distance is a metric and the total weights of two signatures are equal, then EMD is a true metric where non-negativity, symmetry and triangular inequality holds. This allows embedding the signatures into a metric space.

4.2 Constraining EMD flows through Spatial Relation and Negative Cues

EMD, nevertheless, is spatially unconstrained because regions are considered separately during flow. This situation indeed risks random flows particularly when the underlying features are not perfectly robust. Ideally, a fabric of EMD flows should possess the following properties: (i) The flows among common patterns are closed and dense, leading to less association between actual pattern and background. (ii) The flows among common patterns are characterized by high similarity value. (iii) The flows among 'common pattern-background' and 'background-background' patterns are characterized by low similarity value. The correct prioritization of flows to satisfy the above requirements can be enhanced by taking into account spatial information during EMD matching. Tagging spatial information increases the discriminating power of the underlying features. However, common backgrounds, violating the third requirement, cannot be solved by powerful features alone. Such scenarios can be effectively handled by using cues from negative images to repress the undesirable patterns.

Both the spatial information and negative cues can be introduced into EMD matching through the distance measure $dist(i, j)$, or conversely the similarity measure $sim(i, j)$ where $sim(i, j) = 1 - dist(i, j)$. The spatial similarity is measured by considering the neighborhood consistency when matching two regions, formulated as a weighted combination of unary color similarity h_u and the neighborhood similarity h_n. A significance value S is further attached to penalize regions with high proximity to the regions in negative images. The use of negative images to measure the importance of each region is optional but practically useful. Given regions i and j from two signatures and the bag of negative regions $\mathbf{R}^-$ from all negative images, the similarity measure sim can be formulated as

$$sim(i, j) = \min_{p=i,j} S(p) \times h(i, j) \tag{4}$$

where regions suspected as background patterns are assigned lower S. The region-pair similarity is given by

$$h(i, j) = \alpha\, h_u(i, j) + (1 - \alpha)\, h_n(i, j) \tag{5}$$

and the significance value is formulated as

$$S(p) = \begin{cases} 1 - \max h(p,r) & \text{if } \max h(p,r) > T,\ \forall r \in \mathbf{R}^- \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

The parameter α is to weight the importance of color and spatial similarities, while T is to gate the ownership of a candidate region p. Both parameters are set empirically. In order not to over-emphasize spatial constraint, we give higher weight (α=0.6) to color similarity. For the parameter T, a higher value is preferred (T=0.9) in order to exclude regions with high similarity to negative images.

The measure h_u is based on the Euclidean distance in the CIE-L*a*b color space. CIE-Lab color space is expressly designed so that short Euclidean distances correlate strongly with human color discrimination performance. The measure h_u is

$$h_u(i,j) = exp\left[-(\frac{D(i,j)}{\gamma})^2\right] \tag{7}$$

where D is the Euclidean distance of the colors in the CIE-L*a*b color space given by

$$D(i,j) = [(\triangle L)^2 + (\triangle a)^2 + (\triangle b)^2]^{\frac{1}{2}} \tag{8}$$

$h_u(i,j)$ is an exponential function with the steepness of the slope governed by γ, where two regions are regarded as totally non-similar and therefore can be ignored when their distance exceeds 2γ. For our purpose, we set γ to 30.

The neighborhood similarity, or more specifically spatial similarity, of two regions is measured by matching their neighbors in the image space. For the region pair (i,j), we build a weighted bipartite graph (WBG), $G_{i,j} =< U, V, E >$, with U as a set containing the adjacent regions of region i, and V for region j. The weight w_{rs} of the edge e_{rs} connecting the neighbor region $u_r \in U$ and $v_r \in V$ represents the color similarity between the two regions, defined by Equation 7. The neighborhood similarity h_n is measured by performing a maximum weighted bipartite matching [22] between the sets U and V. The similarity, based upon two sets of neighbors, is measured as

$$h_n(i,j) = \sum_{e_{r,s} \in M_{U,V}} w_{rs} \tag{9}$$

where $M_{U,V}$ is the set of edges from the maximum matching of the weighted bipartite graph built from U and V.

5 Local Flow Maximization

The matching patterns of EMD are dispersed and unregulated in nature, resulting in difficulties to expose the common pattern and identify their transformation parameters. Finding common pattern by examining each correspondence separately is susceptible to noise artifacts. This motivates the proposal of a novel predictive approach where a candidate common pattern is initially extracted from each positive image and then matched as a whole iteratively. As such, CPD aims to find the optimum candidate common patterns of positive images that maximize the generic function $\mathcal{H}$ in Equation 1. Matching, when analyzed at such granularity, is more tolerant to noise since the goodness of such matching is determined collectively over a set of correspondences. Besides, when matching is localized, many distracting and irrelevant correspondences have been discarded, resulting in more concentrated and precise analysis.

The candidate common pattern is represented by its centroid and scale which are unknown and have to be discovered to maximize $\mathcal{H}$. One possible solution is through multiple instance learning (MIL). A huge pool of candidate areas with varying scale and centroid values are extracted and projected into a signature space that endows the EMD distance metric. Then, an MIL algorithm can be employed to learn the best instance as common pattern representation [23]. Nonetheless, such scheme is neither efficient nor effective with the presence of noise artifacts. Instead, the proposed predictive approach adopts heuristic search without generating a huge set of candidate areas.

We propose an algorithm, namely Local Flow Maximization (LFM), to find common pattern by analyzing the EMD flows under the expectation-maximization (EM) framework. Given a set of N positive images $\mathbf{I} = \{I_i\}_{i=1}^{N}$, the set of candidate common patterns $\mathbf{P} = \{\mathcal{P}_i\}_{i=1}^{N}$ characterized by both the centroid $\mathbf{c} = \{c_i\}_{i=1}^{N}$ and scale $\mathbf{s} = \{s_i\}_{i=1}^{N}$, is treated as the missing parameter. The set of centroids $\mathbf{c}$ is hidden but can be inferred from the EMD flows $\mathbf{F}^{(n)}$ during iteration n. In the expectation step, the collaborative mining of patterns is performed through localized matching. For each image, the EMD flows between each candidate P_i of image I_i and $P_{i\neq j}$ of other images are analyzed for the estimation of c_i. In the maximization step, the refined value of the centroid c_i is used to redefine the scale s_i for the complete description of $\mathcal{P}_i$ that maximizes the CPD similarity function $\mathcal{H}$. In deriving $\mathcal{H}$ which comes in the form of a maximum likelihood function, two assumptions are made. First, images are assumed independent given the presence of a common pattern, denoted by $\mathcal{C}$. Second, the distribution of $\mathbf{P}$ and $\mathbf{c}$ are assumed to follow an uniform distribution. By the standard EM formulation and Bayes theorem without loss of generality, $\mathcal{H}$ is formulated as

$$
\begin{aligned}
\mathcal{H}(\mathbf{P}|\mathbf{P}^{(n)}) &= E_{\mathbf{c}}\{\log p(\mathbf{c},\mathbf{I}|\mathbf{P},\mathcal{C})|\mathbf{P}^{(n)},\mathbf{I},\mathcal{C}\} \\
&= \sum_{\mathbf{c}} p(\mathbf{c}|\mathbf{P}^{(n)},\mathbf{I},\mathcal{C}) \log p(\mathbf{c},\mathbf{I}|\mathbf{P},\mathcal{C}) \\
&= \sum_{\mathbf{c}} p(\mathbf{c}|\mathbf{P}^{(n)},\mathbf{I},\mathcal{C}) \log \prod_{i=1}^{N} p(c_i,I_i|\mathcal{P}_i,\mathcal{C}) \\
&= \sum_{c_i} p(\mathbf{c}|\mathbf{P}^{(n)},\mathbf{I},\mathcal{C}) \sum_{i=1}^{N} \log \frac{p(\mathcal{P}_i|c_i,I_i,\mathcal{C})p(c_i,I_i|\mathcal{C})}{p(\mathcal{P}_i|\mathcal{C})} \\
&= \sum_{c_i} p(\mathbf{c}|\mathbf{P}^{(n)},\mathbf{I},\mathcal{C}) \sum_{i=1}^{N} \log \frac{p(\mathcal{P}_i|c_i,I_i,\mathcal{C})p(c_i|\mathcal{C})p(I_i|\mathcal{C})}{p(\mathcal{P}_i|\mathcal{C})} \\
&\sim \sum_{c_i} p(\mathbf{c}|\mathbf{P}^{(n)},\mathbf{I},\mathcal{C}) \sum_{i=1}^{N} \log p(\mathcal{P}_i|c_i,I_i,\mathcal{C}) \\
&= \sum_{c_i} \sum_{i}^{N} \delta(c_i|\mathcal{P}_i^{(n)},I_i,\mathcal{C}) \log p(\mathcal{P}_i|c_i,I_i,\mathcal{C}) \\
&= \sum_{i}^{N} \sum_{c_i} \delta(c_i|\mathcal{P}_i^{(n)},I_i,\mathcal{C}) \log p(\mathcal{P}_i|c_i,I_i,\mathcal{C}) \\
&= \sum_{i}^{N} \delta(\hat{c}_i|\mathcal{P}_i^{(n)},\mathcal{C},I_i) \log p(\mathcal{P}_i|c_i,I_i,\mathcal{C}) \qquad (10)
\end{aligned}
$$

where $\hat{c}_i$ is assigned to a single location in the image space through delta function δ. The prior probabilities $p(\mathcal{P}_i|\mathcal{C})$ could be used if c_i and s_i of I_i is known. In Equation 10, the maximization step finds the best candidate common pattern $\mathcal{P}_i$ that optimizes the likelihood function $\mathcal{H}$ through the conditional probability $p(\mathcal{P}_i|c_i,F_i^{(n)},\mathcal{C})$ given $\hat{c}_i$. The expectation step in turn estimates the new value of the position $\hat{c}_i$ from $\mathcal{P}_i^{(n)}$. The algorithm iterates between the two steps until $\mathcal{H}$ converges where $\mathcal{H}^{(n+1)} - \mathcal{H}^{(n)} < \epsilon$. The EM algorithm is guaranteed to converge [24,25] as long as $\mathbf{P}^*$ is chosen such that $\mathcal{H}(\mathbf{P}^*|\mathbf{P}^{(n)}) > \mathcal{H}(\mathbf{P}^{(n)}|\mathbf{P}^{(n)})$ although it might not necessarily be a global maximum.

LFM can be viewed as a 2-class clustering problem where region units of each image are collaboratively clustered into the common pattern or background class. In this aspect, it is similar to the K-Means algorithm. Regions enclosed by a candidate common pattern correlates to the common pattern and vice versa. Similarly, during each iteration, the ownership of each region is iteratively refined until $\mathcal{H}$ is optimized through a variant of the expectation-maximization algorithm. However, K-means operates primarily on feature space while LFM operates on image space and additionally draws hints from matchings. In LFM, there are no class centers upon which a distance measure facilitates assignments. Instead, the boundaries of the common pattern as a localized area are adjusted by looking into the tendency of the EMD flows. While the square mean error (SSE) for K-Means is determined based

on O2O correspondences, $\mathcal{H}$ is determined collectively from the fabric of M2M correspondences.

5.1 Maximization

In the maximization step, the set of candidate common patterns $\mathbf{P}$ from positive images that optimizes the maximum likelihood $\mathcal{H}$ is determined given the estimated $\hat{\mathbf{c}}$ from the expectation step. From Equation 10, this step is expressed as

$$\begin{aligned}\mathbf{P} &= \arg\max_{\mathbf{P}} \ \mathcal{H}(\mathbf{P}|\mathbf{P}^{(n)}) \\ &= \arg\max_{\mathbf{P}} \ \sum_{i}^{N} \log p(\mathcal{P}_i|c_i, I_i, \mathcal{C}) \end{aligned} \tag{11}$$

where the conditional probability $p(\mathcal{P}_i|c_i, I_i, \mathcal{C})$ that contributes towards $\mathcal{H}$ can now be defined in terms of pair-wise localized matching as follows

$$p(\mathcal{P}_i|c_i, I_i, \mathcal{C}) = \prod_{j=1, j\neq i}^{N} \mathrm{EMD}(\mathcal{P}_i, \mathcal{P}_j) \tag{12}$$

Clearly, getting the optimum set of $\mathcal{P}_i$ is intractable and we have to resort to non-linear optimization algorithms such as gradient ascent where additional constraints are imposed on $\mathcal{P}_i$. The constraints include the minimum number of pixels and the permissible area localized to $\hat{c}_i$ over which $\mathcal{P}_i$ could take shape. However, using gradient ascent is computationally expensive because a larger amount of EMD matching is involved during each hill-climbing iteration. To improve speed efficiency, the localization of $\mathcal{P}_i$ is performed on a discrete set of candidate areas and the best candidate is selected as elaborated below.

The set of candidate common patterns that maximizes the conditional probability in Equation 11 does not always result in a semantically optimum scale because of an inherent bias towards signatures composed of few segmented regions. In fact, similarity might reach its maximal value at a trivial situation where a signature consists of only one region representing a common pattern. Weighted fusion or thresholding can be employed to handle the trade-off between weight and scale. Nonetheless, such schemes require heuristic settings and undermine robustness. Instead, given an assorted set of scales in an input image, the maxima of the first-order derivatives of the similarity values from cross-matchings with other candidate common patterns is picked as the optimum scale. This is coupled with a monotonic constraint on the similarity value to preserve convergence.

For pragmatic reasons, a bounding box centered on $\hat{c}_i$ is used to represent $\mathcal{P}_i$. Coverage can be improved at the expense of speed by employing more robust shapes such as ellipse that offers a larger degree of freedom in terms of pose and orientation. Given a positive image, a set of bounding box, ordered by increasing widths and centered on the centroid $\hat{c}_i$, is generated. The candidate common pattern is matched with candidates of all other positive images and then averaged. A prominent drop in the similarity value while walking through the scale is evident as shown in Figure 4 when the boundary box outgrows the visually coherent common pattern, thus providing a robust choice for scale selection.

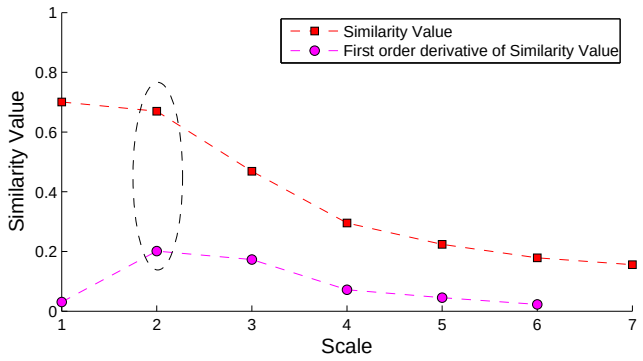


Fig. 4. The matching similarity value, when $\mathcal{P}_i$ is varied over a range of sorted scales, experiences a prominent drop in similarity when the scale outgrows the common pattern as highlighted by the ellipse.

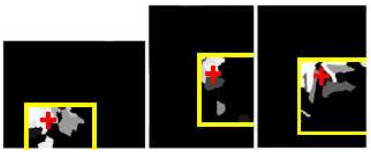
5.2 Expectation

The loose embedding of spatial information into EMD matching encourages the 'crowding' of high similarity flows at the spot that exhibits both structural and photometric resemblance among the candidate common patterns. Realigning the centroid of the candidate common pattern towards such spot given the candidate common pattern $\mathcal{P}_i^{(n)}$ improves the probability density of the flows, and thus leads towards the discovery of an exact common pattern. The crowding spot $\hat{c}_i$ is a single point in the image space expressed through $\delta(\hat{c}_i|\mathcal{P}_i^{(n)}, \mathcal{C}, I_i)$ in Equation 10.

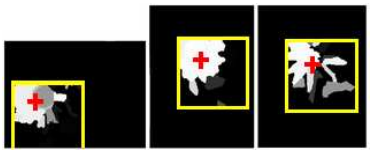
Given the *optimum* candidate common pattern $\mathcal{P}_i$ from the maximization step, the flows from the localized EMD matchings of these areas are collected, and a refined value of $\hat{c}_i$ for each image is computed as the weighted mean of each location in the sub-image space. For each positive image I_i, the collection of flows $\mathbf{F}_i$ having I_i as the destination image, is extracted from the set of pairwise EMD matchings of the optimum candidate common patterns, readily available

(a) Original Images

(b) Confidence Images



(c) Iteration 1



(d) Iteration 2

(e) Iteration 3

(f) Iteration 4

Fig. 5. Fast-tracking of $\hat{c}_i$ to the best optimum position during the expectation step. The scale value $s_i^{(n)}$ remains the same during the whole process. The confidence image shows the weight η of of all regions. The red cross shows the position of $\hat{c}_i$.

from the previous maximization step. Each pixel location x in the sub-image space is weighted by accumulating the similarity of the flows to the region that the pixel belongs to. Denoting $\mathbf{F}_{i,x}$ as the subset of flows that streams into the region that the pixel x holds ownership, and $sim(f)$ as the similarity value of a flow f, the new centroid $\hat{c}_i$ is formulated as the weighted mean of each pixel as follows

$$\hat{c}_i = \frac{\sum_{x \in \mathcal{P}_i} x \times \eta(x)}{\sum_{x \in \mathcal{P}_i} \eta(x)} \tag{13}$$

where

$$\eta(x) = \sum_{f \in \mathbf{F}_{i,x}} sim(f) \tag{14}$$

Inspired by the iterative closest points (ICP) algorithm in [26] and the flow transformation (FT) algorithm in [27], $\hat{c}_i$ is computed repeatedly using the same scale in the expectation step, each time realigning the center of the bounding box to the centroid value. Convergence is preserved through the constraint that likelihood function increases monotonically during each iteration. As a result, $\hat{c}_i$ is pushed to the best possible location as shown in Figure 5. This process is referred to as the *fast-tracking* of $\hat{c}_i$ because it optimizes speed by reducing the number of EM iterations.

5.3 Initial Prediction of Common Pattern

As in conventional EM algorithms, LFM is sensitive to initialization. In LFM, the initialization of the centroid $c_i^{(0)}$ of the set of common patterns $\mathcal{P}_i$ is obtained by analyzing the density of EMD flows conducted globally at the image level. The EMD flows, although noisy, provide the initial cue for locating common pattern. Given a set of R regions $\{r_i\}_{i=1}^{R}$ in an image, the density probability $p_r(x)$ of the pixel x can be estimated through the extrapolation of the regions by means of weighted Parzen-window. Each region r_i is represented by its centroid and weighted by a confidence value $\hat{\eta}(r_i)$. Using the same methodology in the computation of $\eta(r_i)$ in Equation 14, $\hat{\eta}(r_i)$ is determined by accumulating the similarity of EMD flows to r_i and then normalized over all other samples. Thus, $c_i^{(0)}$ can be assigned based on a maximum a posterior criterion as follows

$$c_i^{(0)} = \arg\max_x p_r(x) \tag{15}$$

where the Parzen-window density estimate is given by

$$p_r(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\eta}(r_i)}{h_r^d}\varphi(\frac{x - x_i}{h_r}) \tag{16}$$

and h_r is the size of the window width parameter which determines the smoothness of the density function. A Gaussian kernel window is used and φ is formulated as

$$\varphi(\tilde{x}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\tilde{x}^2}{2}} \tag{17}$$

As a result, the areas with high concentration of heavily weighted flows would experience higher density value, and using the location with the maximum density provides a good estimate of the initial centroid $c_i^{(0)}$.

6 Result and Experiments

We conduct both qualitative and quantitative assessment to verify the performance of the proposed approach. We use F-measure [28,6] as the metric for performance evaluation. F-measure is a popular measure used in Information Retrieval (IR) as the weighted harmonic mean of precision and recall. For CPD, precision measures the accuracy of detection, while recall measures the ability of completely locating a common pattern. F-measure combines both measures to provide only a single value that effectively quantify the quality of the detected common pattern. Denoting $\mathcal{G}$ as the groundtruth pattern and $\mathcal{D}$ as the detected pattern, F-measure is defined as

$$\text{F-measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{18}$$

where

$$\text{recall} = \frac{\text{area}(\mathcal{G} \cap \mathcal{D})}{\text{area}(\mathcal{G})} \tag{19}$$

$$\text{precision} = \frac{\text{area}(\mathcal{G} \cap \mathcal{D})}{\text{area}(\mathcal{D})} \tag{20}$$

In the experiments, a total of 14 common patterns is used for testing. For each pattern, three positive images and one negative image are given for CPD. To assess the robustness of proposed approach, large variations of location, scale, rotation and viewpoint are introduced to common patterns during image collection. To highlight the effect of noise to CPD, all images are shot under complex background setting.

6.1 Performance Comparison

To assess the performance, we compare three approaches: MWBG [10], IMCS [5], and our approach, namely LFM. MWBG employs one-to-one bipartite graph matching to find correspondences between two sets of block uniformly and densely extracted from images. Procrustes analysis is then performed to estimate optimal transformation based on block correspondence. The matching and transformation steps are iterated until $\mathcal{H}$ in Equation 1 optimizes. IMCS operates on region units and poses CPD as a subgraph isomorphism problem. The backtrack depth first search, a brute-force technique, is utilized to mine maximum common subgraph from the ARG representations of images. Table 1 briefly summarizes the three compared approaches. For IMCS and LFM, the tool in [29] is employed for image segmentation. For LFM, each region is set to a minimum of 100 pixels, resulting in a range of 50 to 300 region units. For IMCS, due to brute-force search, the setting has to be deliberately tuned to less than 90 regions to achieve a reasonable speed.

Figure 6 presents the quantitative performance of different approaches. Overall, LFM outperforms both MWBG and IMCS in F-Measure with an average score of 0.66 compared to 0.30 and 0.34 respectively, indicating the advantages of using M2M coupled with a localized matching strategy over conventional O2O matching techniques. M2M successfully exhibits dense matching where the erroneously segmented regions can still be correctly matched with EMD. IMCS, in contrast, employs O2O mapping and results in sparse matching where the broken regions are left alone without correspondences. Figure 7 contrasts the matching robustness of M2M and O2O. Through M2M matching strategy, LFM is more tolerant to segmentation error and capable of matching fragmented regions. MWBG, on the other hand, operates on block units

Table 1
Summary of the CPD algorithms used in our experiment

Description	LFM	IMCS	MWBG
Matching Unit	Region	Region	Block (Point)
Matching Technique	Localized (M2M)	Graph (O2O)	P2P (O2O)
Features	Color	Color	Color
Representation	Weighted point set	ARG	Bipartite graph
Segmentation Tolerance	Yes	No	Not applicable
Negative Image	Yes	No	No
Search Complexity	Non-exhaustive	Exhaustive	Non-Exhaustive

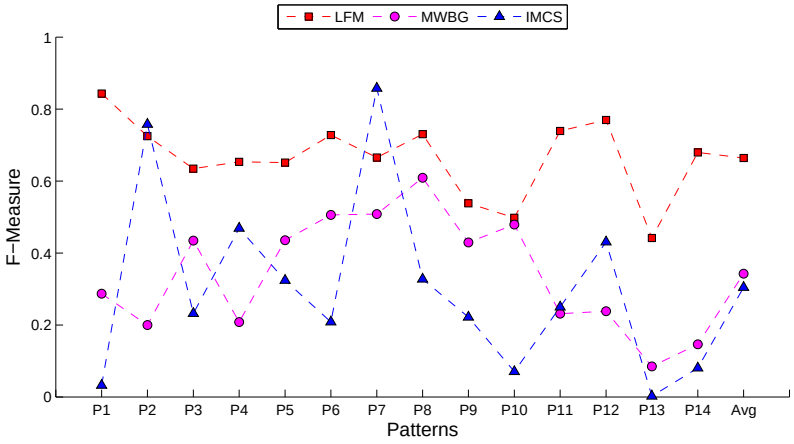


Fig. 6. F-Measure performance among LFM, MWBG and IMCS.

and therefore is inherently sensitive to scale. From our observation, when procrustes analysis is performed on correspondence pairs obtained through bipartite graph matching, noise artefacts skew the computation of the transformation parameters. Such errors propagate through the iterations and cause ineffective mining.

Image segmentation often requires heuristic settings such as the minimum size of a region. Figure 8 shows the performance of different approaches against the "minimum size" heuristic. To highlight the effects of the heuristic, the patterns are matched without interference from other artefacts especially large scale variation and background clutter. The performance of IMCS is poor when regions are over-segmented into ambiguous pieces. In contast, LFM yields an almost constant F-measure, which indicates the resilience of M2M matching

(a) Original segmented images

(b) Common pattern by IMCS

(c) Common pattern by LFM

Fig. 7. Common patterns extracted by (b) IMCS and (c) LFM contrast the matching robustness of O2O and M2M. In IMCS, O2O results in sparse matching of erroneously segmented regions. LFM effectively utilizes EMD to densely match these regions.

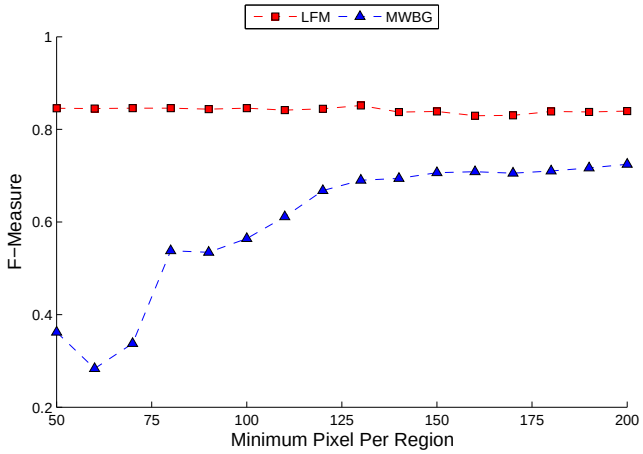


Fig. 8. Sensitivity of different approaches when the constraint "minimum pixel per region" is imposed during segmentation.

towards structural inconsistencies arising from segmentation. Figure 9 further shows an example to contrast LFM and IMCS under three different settings.

To illustrate the effectiveness of LFM, Figure 10 shows the ability of LFM in locating common pattern in three iterations. The initial value of $\mathbf{c}^{(0)}$ is

(a) Input Images

(b) Common pattern by IMCS

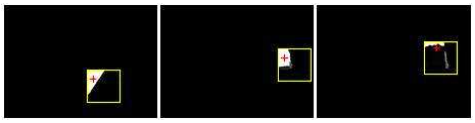
(c) Common pattern by LFM

Fig. 9. The mined patterns by (b) IMCS and (c) LFM when the minimum size of segmented regions is set to 50 (left), 100 (center) and 200 (right) pixels respectively.

manually placed in order to emphasize the power of LFM even when the initialization encloses only a small portion of the common pattern. The location of $\mathbf{c}^{(n)}$ is marked by the cross (+) while the box shows the candidate common pattern $\mathbf{P}^{(n)}$. A new candidate common pattern is selected at the start of each iteration in the maximization step. Then, $\mathbf{c}^{(n)}$ is determined and iteratively recomputed to push $\mathbf{P}^{(n)}$ to the most optimum location in the expectation step. The resting positions of $\mathbf{c}^{(n)}$ during each iteration are shown in the right columns of the figure. In this experiment, we could see that $\mathbf{c}^{(n)}$ successfully converges to the common pattern within 3 iterations, accompanied by an automatic readjustment in the scale of the candidate common pattern.

Figures 11 and 12 show the robustness of LFM towards rotation and viewpoint difference respectively under background clutter. LFM successfully locates the common patterns despite the challenges. The spatial constraint we introduce in EMD indeed enforces the orderly matching of regions, which led to more robust localization and EM framework. However, the limitations of using bounding box to represent common pattern is prominent when finding patterns such as P5 under viewpoint variation in Figure 12(b). Intuitively, the performance can be improved if other representations such circle or ellipse is used, with the expense of computational cost. Figure 13 further shows the effectiveness of LFM towards large scale changes, in addition to the rotation and viewpoint variation. LFM is able to locate the common pattern, but with relatively lower

(a) Input images



(b) Start of iteration 1

(c) End of iteration 1

(d) Start of iteration 2

(e) End of iteration 2

(f) Start of iteration 3

(g) End of iteration 3

(h) CPD result

Fig. 10. Intermediate output of LFM iterations. The left column shows the selected common pattern **P** during the expectation step while the right column shows its resting position from the fast-tracking cycles in the maximization step.

recall for pattern with larger scale. Figure 14 shows the average F-measure of LFM on the tested patterns when one of the positive images is transformed with a predefined range of scales. LFM is relatively stable across various scale changes with F-measure of approximately 0.6.

Figure 15 shows the robustness of LFM towards patterns with slight changes in color under different lighting conditions, from the brightest setting in the second image to the darkest setting in the fourth image. The common pattern can be effectively mined as long as the variation is not too dramatic. Severe variation in illumination impairs dense region matching, and the use of sparse features such as corners or edges is more appropriate in these cases. Figure 16 shows some challenging examples where the patterns are hidden in highly cluttered backgrounds, and further obscured by scale, viewpoint and rotation. LFM successfully highlights the common pattern despite such demanding environments. Some less desirable results are shown in Figure 17. In pattern P13, an erroneous pattern is detected where the pattern in the second image is trapped in a local maximum, while in P14, an inflated version of the common pattern is detected. Apparently, LFM is sensitive to initialization and

(a) Pattern P1

(b) Pattern P2

(c) Pattern P3

Fig. 11. Finding common patterns under rotations.

(a) Pattern P4

(b) Pattern P5

(c) Pattern P6

Fig. 12. Finding common patterns under viewpoint variation.

it is observed that the erroneous pattern in P13 is still visually similar to the other two common patterns in terms of low level features.

Further evaluations are conducted with the common patterns subjected to the combination of all noise. The patterns are obscured by rotation, scale, skew and viewpoint with heavy background clutter. Evaluation is performed qualitatively through manual observation. CPD manage to consistently identify the pattern-of-interest despite the difficulties as shown in Figure 18.

To assess the role of negative images, we repeat LFM by using only positive

(a) Pattern P7

(b) Pattern P8

Fig. 13. Finding common patterns under scale changes, in addition to rotation and viewpoint variation.

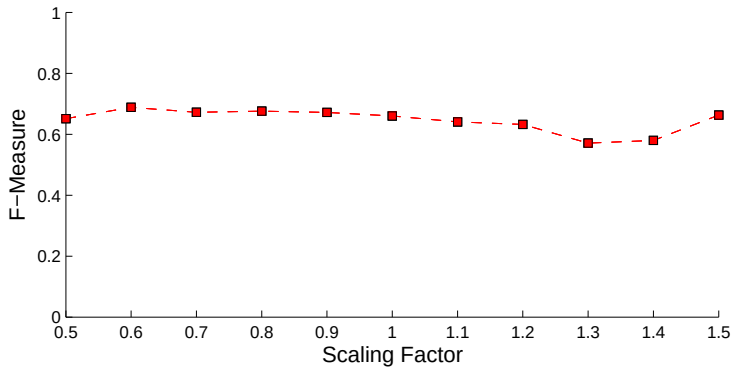


Fig. 14. F-Measure of patterns when transformed over a range of scaling factor

(a) Pattern P9

Fig. 15. Finding common patterns under lighting variation.

images. The results show that the F-measure drops from 0.66 to 0.57. It is observed that the impact towards recall is random but there is a consistent drop in precision for almost all patterns. Indeed, negative images are passive cues that do not assert any patterns as the common pattern, but rather rule out irrelevant patterns through suppression of noise artefacts, resulting in a more precise common pattern. However, negative images could be counter-productive if inappropriately used, as observed in the evaluations on patterns P13 and P14 where the performance of F-Measure drops by 0.28 and 0.15. Such scenarios happen when negative images surpress the common pattern instead of the intended background artefacts.

(a) Pattern P11

(b) Pattern P12

Fig. 16. Finding common patterns superimposed on highly cluttered background.

(a) Pattern P13

(b) Pattern P14

Fig. 17. Results with erroeneous and sub-optimal patterns using LFM.

6.2 Speed Efficiency

The speed of LFM and IMCS is impacted by the number of regions. Table 2 shows their speed with different region settings. When the number of regions is large, LFM is significantly faster than IMCS, but slightly slower than MWBG which uses fixed amount of blocks rather than regions. LFM is less sensitive to the increase of regions compared to IMCS.

Table 2
Average speed (seconds) for the pattern P3 using different segmentation settings

Minimum #pixel per region	150	175	200	300	400
Total number of regions	329	297	270	256	162
LFM	92.77	85.02	80.94	79.49	93.03
IMCS	2521.79	400.97	103.47	14.25	6.2
MWBG	72.63				

The proposed LFM is essentially an efficient polynomial time algorithm. If B

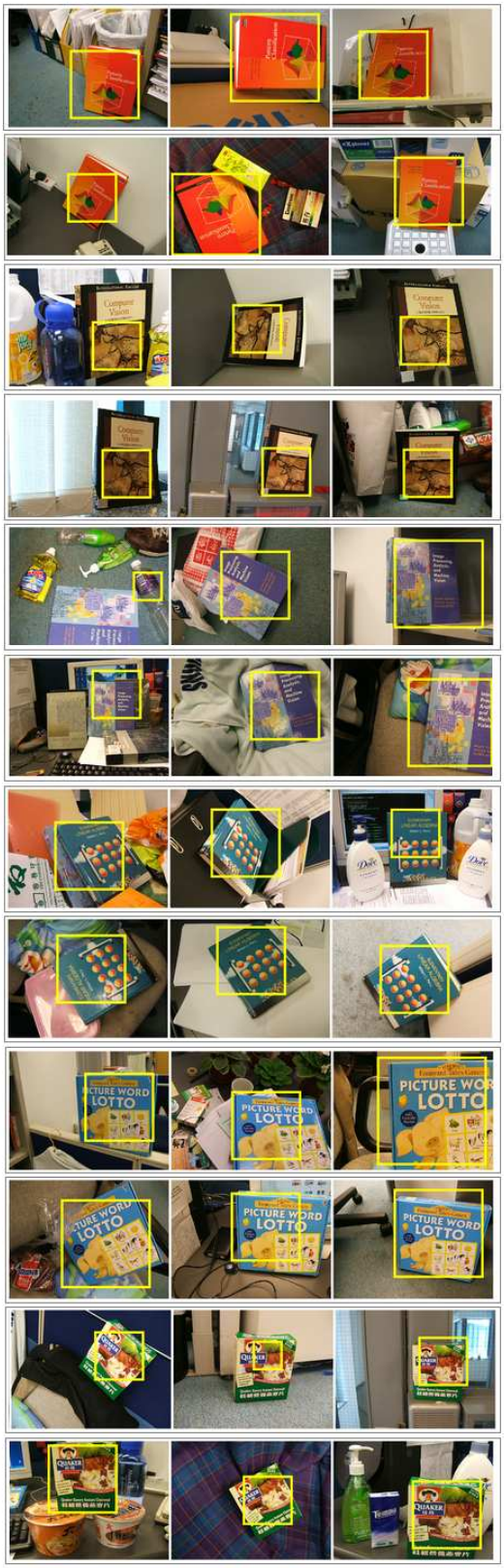


Fig. 18. CPD results on patterns with a variety of image noise

iterations are required for LFM, the total running time is $O(T_i + B(T_m + T_e))$, where T_i, T_m and T_e are the times spent for initialization, the maximization and expectation step, respectively. These times are linearly dependent on the EMD matching. Each EMD matching requires $O(r^3 \times \log(r))$ time where r is the number of regions in the signatures [19]. Given N images, initialization performs matching for $N_p = \frac{N(N-1)}{2}$ times over all image-pairs. The expectation step is slower since image-pair matchings are repeated A times for the fast-tracking iterations. The maximization step incurs the most computational cost since image-pair matchings are repeated over all scale variations, resulting in a $S_p = \frac{s(s-1)}{2}$ fold increase in computational time when a total of s scales is used for analysis. In our experiments, s is 7, N is 3, r ranges from 30 to 200, A from 2 to 20 and B from 3 to 6. The fast-tracking enhancement proposed in the expectation step has been able to alleviate the cost by minimizing the number of required iterations.

The running time of MWBG depends on the graph matching algorithm which takes $O(CBb_i^3)$, where b_i is the number of blocks in each image and B is the number of iterations required for each run. The maximum weighted matching has a time complexity of $O(b_i^3)$. For robustness purposes, MWBG is repeated using C different transformation initializations. On the contrary, the speed of IMCS increases exponentially with the number of regions owing to brute-force search. The time complexity for IMCS is $O(2 \times r_i!)$ where r_i is the number of regions in the images.

6.3 Limitations and Future Directions

Since we do not consider variations such as rotation and stretching when performing matching, our algorithm is unable to extract the full set of affine parameters which might be useful for certain applications such as stereo calibration. However, since our algorithm manages to produce a good localization of the common patterns as demonstrated in our experiments, the affine parameters estimation can be carried out as a separate post-processing step in a more concise manner on the mined patterns.

Due to the employment of color as the underlying feature for matching, our approach is inevitably limited to objects with multiple regions. To be successful on a different category of images, where the common pattern is encapsulated within a single region unit with prominent shape or visual point details, it is imperative to extend LFM to sparse matching on point features. Recently, the local features based on keypoints [2,30,31] are shown to be powerful and discriminative for a wide range of vision-related tasks. Coupled with sparse matching techniques such as the Hungarian algorithm [22], Integer Quadratic Programming matching [8] and One-to-One symmetric (OOS) matching [32] in

place of EMD as the underlying matching tool, these features can be employed by LFM to handle this class of images as well.

Finally, the current speed of LFM still cannot efficiently handle a large number of input images. It is interesting to explore how fast the variants of EMD matching algorithms such as embedded EMD [33] can be employed to accelerate pattern matching.

7 Application of CPD to Image Retrieval

CPD can be exploited for relevancy feedback in retrieval, by collecting the positive and negative labels of few images from users to refine search. The learning task of relevance feedback [34–36] is normally tedious and needs to be repeated for many rounds to arrive at a satisfactory result. With CPD, relevance feedback is composed of three simple steps – coarse retrieval, CPD and fine retrieval. The coarse retrieval performs an initial search through conventional query-by-image search, for instance. The retrieved relevant images are marked and then CPD is employed to mine the common pattern. The mined common pattern is subsequently used as "query-by-pattern" for fine retrieval. Query-by-pattern is a lazy process which commences only upon a request for retrieval. In contrast, traditional retrieval systems are generally more active where comprehensive models or indexing systems have to be put in place for retrieval to be effective. The new paradigm focuses on engaging more interaction at contact time by encouraging users to supply more examples. CPD is useful in this respect to highlight the important patterns for query. Indeed, query-by-pattern is interesting for its potential to improve retrieval results, even on an unprocessed database.

To demonstrate the effectiveness of CPD in retrieval, we use a database composed of 1068 images with 14 common patterns. The common patterns are shot under different background clutters, at varying viewpoints, rotation, scale and lighting changes. To increase the diversity of the database, 200 random images which do not have common patterns are added. For performance evaluation, we compare four different approaches: (1) CPD with proposed LFM (2) DD [12] and (3) EM-DD [13] and (4) Histogram intersection [37]. DD and EM-DD are classical algorithms in multiple instance learning (MIL). In both algorithms, each image is divided into overlapping blocks of three separate sizes (15x15, 30x30, 40x40), which form the instances of a bag. DD and EM-DD find a feature point which is common in positive bags while rare in negative bags. The common pattern, embedded in a feature space, is mined through gradient ascent. The search is repeated using multiple starting points to improve robustness. As a baseline, histogram intersection using query-by-image is used to justify the performance improvement of our approach.

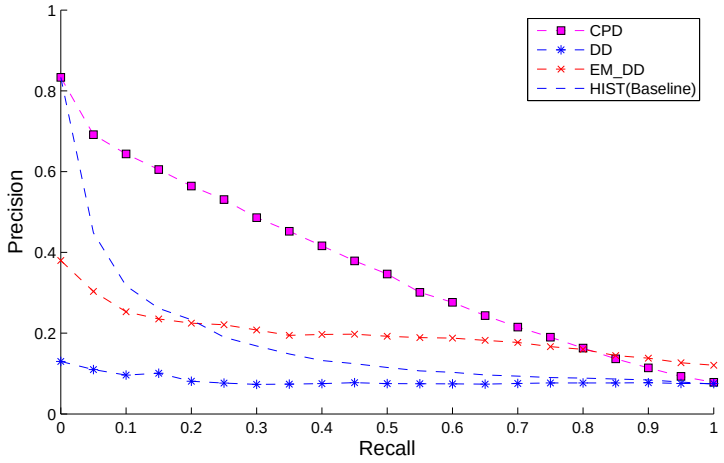


Fig. 19. Precision-recall performance of five different approaches.

In our approach CPD, the similarity between a mined common pattern and an image is performed through histogram intersection during fine retrieval. For DD and EMD-DD, we use the techniques proposed in [14] to perform query-by-pattern. Basically, the located feature point, which represents the common pattern, is used as the feature vector to retrieve images. In the experiments, CPD uses 3 positive and 1 negative images for pattern mining. DD and EM-DD, on the other hand, require more bags in order to be precise. As a results, 10 positive and 20 negative images are used.

Figure 19 shows the recall-precision curve of four different approaches, averaged over 56 queries involving all fourteen patterns. Through experiments, we demonstrate that CPD successfully retrieves common patterns embedded in various backgrounds despite the high degree of variations in scale and rotation in the database. A significant improvement over the baseline is observed indicating the advantages of using pattern as the query for retrieval. Patterns are more powerful in representing the semantic content of the images, as keywords in representing text documents. Therefore, it is better positioned to capture the intention of the users for searching. DD and EM-DD are not as good as CPD in general because the mined pattern is less perfect compared to CPD. Figure 20 shows the common pattern mined by CPD and DD respectively.

8 Conclusion

We have presented our approach for common pattern mining in multiple images. Several critical issues on matching including feature robustness, matching robustness and noise artifact are discussed. We propose M2M with the aid of EMD as the sensible matching technique when image-segmented regions are used. To handle the unregulated matching patterns in M2M, we loosely em-

(a)

(b)

Fig. 20. The mined common patterns by (a) CPD and (b) DD. In DD, the best of the hypothesis mined at different initializations is selected. In (b), five best hypothesis are shown.

bed the spatial information of region into the EMD similarity measure and further propose the LFM framework. LFM adopts localized matching where the candidate common pattern of each image is extracted and matched locally across multiple images. EM is used to iteratively refine the candidate common pattern until the optimum patterns are mined. Experimental results show that our proposed LFM is robust to image segmentation. To demonstrate the potential of CPD for image retrieval, we also conduct experiments to contrast retrieval with and without CPD. The experiments show the power of query-by-pattern, especially with the aid of the proposed CPD technique, in overcoming background clutter and various transformations for retrieving the object-of-interest.

Acknowledgements

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 118905).

References

[1] J. Ponce, M. Hebert, C. Schmid, Zisserman (Eds.), Towards Category-Level Object Recognition, Vol. 4170, Springer-Verlag Lecture Notes in Computer Science, 2006.

[2] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[3] P. Hong, T. S. Huang, Inexact spatial pattern mining, in: Workshop on Discrete Mathematics and Data Mining, 2002.

[4] P. Hong, T. S. Huang, Spatial pattern discovery by learning a probabilistic parametric model from multiple attributed relational graph, Discrete Applied Mathematics 139 (1).

[5] H. Jiang, C. W. Ngo, Image mining using inexact maximum common subraph of multiple ARG, in: International Conference on Visual Information System, 2003, pp. 446–449.

[6] S. Todorovic, N. Ahuja, Extracting subimages of an unknown category from a set of images, in: Computer Vision and Pattern Recognition, 2006.

[7] H. K. Tan, C. W. Ngo, Common pattern discovery using earth mover's distance and local flow maximization, in: International Conference on Computer Vision, Vol. 2, 2005, pp. 1222–1229.

[8] A. C. Berg, T. L. Berg, J. Malik, Shape matching and object recognition using low distortion correspondences, in: Computer Vision and Pattern Recognition, Vol. 1, 2005, pp. 26–33.

[9] S. Lazebnik, C. Schmid, J. Ponce, Semi-local affine parts for object recognition, in: British Machine Vision Conference, Vol. 2, 2004, pp. 779–788.

[10] H. Jiang, C. W. Ngo, Graph-based image matching, in: International Conference on Pattern Recognition, Vol. 3, 2004, pp. 658–661.

[11] A. Ratan, O. Maron, T. Lozano-Perez, A framework for learning query concepts in image classification, in: Computer Vision and Pattern Recognition, Vol. 1, 1999, pp. 423–429.

[12] O. Maron, A. L. Ratan, Multiple-instance learning for natural scene classification, in: International Conference on Machine Learning, 1998, pp. 341–349.

[13] Q. Zhang, S. A. Goldman, EM-DD: An improved multiple-instance learning technique, in: Neural Information Processing Systems, 2001, pp. 1073–1080.

[14] Q. Zhang, S. A. Goldman, W. Yu, J. E. Fritts, Content-based image retrieval using multiple-instance learning, in: International Conference on Machine Learning, 2002, pp. 682–689.

[15] Y. Chevaleyre, J. D. Zucker, Solving multiple-instance and multiple-part learning problems with decision trees and rule sets. Application to the mutagenesis problem, Vol. 2056, Lecture Notes in Artificial Intelligence, 2001.

[16] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: Neural Information Processing Systems, 2002, pp. 561–568.

[17] T. Quack, V. Ferrari, B. Leibe, L. V. Gool, Efficient mining of frequent and distinctive feature configurations, in: International Conference on Computer Vision, 2007.

[18] J. Yuan, Y. Wu, Spatial random partition for common visual pattern discovery, in: International Conference on Computer Vision, 2007.

[19] Y. Rubner, C. Tomasi, L. J. Guibas, The earth mover's distance as a metric for image retrieval, International Journal of Computer Vision 20 (2) (2000) 99–121.

[20] Y. Rubner, C. Tomasi, L. J. Guibas, The earth mover's distance, multi-dimensional scaling, and color-based image retrieval, in: Proceedings of the ARPA Image Understanding Workshop, 1997, pp. 661–688.

[21] Y. Rubner, C. Tomasi, L. J. Guibas, Edge, junction and corner detection using color distributions, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (11) (2001) 1281–1295.

[22] J. A. McHugh, Algorithmic graph theory, Prentice Hall, 1990.

[23] C. Yang, M. Dong, F. Fotouhi, Region based image annotation through multiple-instance learning, in: ACM International Conference on Multimedia, 2005.

[24] G. McLachlan, T. Krishnan, The EM algorithm and extensions, John Wiley and Sons, 1997.

[25] R. M. Neal, G. E. Hinton, Learning in Graphical Models, Kluwer Academic Press, 1998, Ch. A view of the EM algorithm that justifies incremental, sparse, and other variants.

[26] P. J. Besl, N. D. McKay, A method for registration of 3-d shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (2) (1992) 239–256.

[27] S. Cohen, L. J. Guibas, The earth mover's distance under transformation sets, in: International Conference on Computer Vision, Vol. 2, 1999, pp. 1076–1083.

[28] M. Steinbach, G. Karypis, V. Kumar, A comparison of document clustering techniques, in: Proc. KDD Workshop on Text Mining, 2000.

[29] P. F. Felzenszwalb, D. Huttenlocher, Image segmentation using local variation, in: Computer Vision and Pattern Recognition, 1998, pp. 98–104.

[30] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, International Journal of Computer Vision 60 (1) (2004) 63–86.

[31] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: British Machine Vision Conference, 2002, pp. 384–393.

[32] W. L. Zhao, C. W. Ngo, H. K. Tan, X. Wu, Scale and affine invariant interest point detectors, Near-Duplicate Keyframe Identification with Interest Point Matching and Pattern Learning 9 (2) (2007) 1037–1048.

[33] K. Grauman, T. Darrell, Fast contour matching using approximate earth mover's distance, in: Computer Vision and Pattern Recognition, 2004.

[34] X. He, W.-Y. Ma, O. King, M. Li, H. Zhang, Learning and inferring a semantic space from user's relevance feedback for image retrieval, in: ACM International Conference on Multimedia, Vol. 1, 2004, pp. 39–48.

[35] T. S. Huang, X. S. Zhou, Image retrieval by relevance feedback: from heuristic weight adjustment to optimal learning methods, in: International Conference on Image Processing, 2001.

[36] Y. Rui, T. S. Huang, A novel relevance feedback technique in image retrieval, in: ACM International Conference on Multimedia, Vol. 3, 1999, pp. 2–5.

[37] M. J. Swain, D. H. Ballard, Indexing via color histogram, in: International Conference on Computer Vision, 1990, pp. 390–393.