รายงานวิจัยฉบับสมบูรณ์

(ภาษาไทย) **การปรับปรุงสมรรถนะของเทคนิคการรู้จำวัตถุที่ใช้แนวทางการเปรียบเทียบคุณสมบัติเชิงลักษณะในบริเวณเฉพาะที่**

(ภาษาอังกฤษ) **On Improving the Performance of Object Recognition with Local Appearance Feature Matching**

ผู้วิจัย: นายประดิษฐ์ มิตราปิยานุรักษ์
ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์
มหาวิทยาลัยศรีนครินทรวิโรฒ
114 สุขุมวิท 23 กรุงเทพฯ 10110
โทรศัพท์ที่ทำงาน 02 649 5000 ต่อ 8615
อีเมลล์ praditm@swu.ac.th

ทุนสนับสนุน: เงินงบประมาณ เงินรายได้คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒประจำปีงบประมาณ 2552 (สัญญาเลขที่ 247/2552)

ชื่อโครงการ: การปรับปรุงสมรรถนะของเทคนิคการรู้จำวัตถุที่ใช้แนวทางการเปรียบเทียบคุณสมบัติเชิงลักษณะในบริเวณเฉพาะที่
ผู้วิจัย : ประดิษฐ์ มิตราปิยานุรักษ์
หน่วยงาน: ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์
มหาวิทยาลัยศรีนครินทรวิโรฒ
แหล่งทุนอุดหนุนการวิจัย: เงินงบประมาณ เงินรายได้คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ประจำปีงบประมาณ 2552 (สัญญาเลขที่ 247/2552)
ปีที่ทำวิจัยเสร็จ : กุมภาพันธ์ พ.ศ. 2555

บทคัดย่อ

การรู้จำวัตถุเชิงมองเห็นได้จากภาพ สามารถนำไปประยุกต์ใช้ได้หลายหลาย เช่น การรู้จำสถานที่ การรู้จำวัตถุเฉพาะ การรู้จำชนิดของวัตถุ เป็นต้น ในปัจจุบันได้มีการพัฒนาอัลกอริทึมโดยใช้แนวทางคุณลักษณะท้องถิ่นสำหรับปัญหานี้ โดยสามารถแบ่งได้เป็น 2 วิธี แบบแรกคือวิธีที่อิงการเปรียบเทียบกลุ่มของคุณลักษณะท้องถิ่น เช่น SIFT, SURF เป็นต้น ซึ่งโดยทั่วไปอัลกอริทึมในกลุ่มนี้จะสามารถรู้จำและระบุตำแหน่งของวัตถุเฉพาะในภาพได้ ส่วนแบบที่สองคือวิธีที่ใช้ Bag-of-Visual-Word ซึ่งพัฒนาต่อมาจากแบบแรกไว้สำหรับการรู้จำที่มีวัตุจำนวนมากในฐานข้อมูล อย่างไรก็ตามวิธีการรู้จำทั้งสองแบบยังมีข้อจำกัด ตัวอย่างเช่น การรู้จำแบบเปรียบเทียบกลุ่มของคุณลักษณะท้องถิ่นจะทำงานได้ดีเมื่อมีจำนวนวัตถุในฐานข้อมูลมีไม่เยอะ ถึงแม้ว่าต่อมาจะมีการพัฒนาวิธีการรู้จำแบบ Bag-Of-Visual-Word มาสำหรับปัญหานี้ มันก็ยังมีข้อจำกัดคือต้องมีวัตถุชิ้นเดียวปรากฏอยู่ในภาพ ซึ่งทำให้การใช้ประยุกต์งานมีข้อจำกัดที่ต้องมีผู้ใช้เข้ามาเกี่ยวข้องในขั้นตอนการรู้จำด้วย

ในงานวิจัยนี้เราจึงต้องการที่จะศึกษาในรายละเอียดถึงข้อจำกัดต่าง ๆของอัลกอริทึมการรู้จำที่ใช้แนวทางคุณลักษณะท้องถิ่น โดยเป้าหมายหลักคือต้องการศึกษาข้อจำกัดและหาแนวทางการปรับปรุงอัลกอริทึมเพื่อแก้ไขข้อจำกัดที่ค้นพบ โดยประเด็นหลักที่เราต้องการศึกษาคือความถูกต้องแม่นยำ ประสิทธภาพ (ความเร็วในการทำงาน) ความยืดหยุ่นต่อจำนวนวัตถุในฐานข้อมูล นอกจากนี้เรายังต้องการศึกษาการประยุกต์ของอัลกอริทึมในสองปัญหาที่แตกต่างแต่สัมพันธ์กัน ได้แก่ การรู้จำวัตถุฉพาะ และ การรู้จำวัตถุตามประเภท

Project title: On Improving the Performance of Object Recognition with Local Appearance Feature Matching

Research : Pradit Mittrapiyanuruk

Affiliation: Mathematics Department, Faculty of Science, Srinakharinwirot University

Sponsor: Srinakharinwirot University 2009 (Grant No.247/2552)

Finished year: February 2012

Abstract

Visual recognition of scene [36] can be applied to various applications e.g. location recognition, specific object recognition, object category recognition. Recently, two major categories of researches based on local feature are developed for this purpose. The first one [3], [16], [19] is based on the constellation of local features (e.g. SIFT, SURF) in which they are proposed to simultaneously recognize and localize objects in the image. For the second category, the recognition algorithms based on Bag-Of-Visual-Words (BOVW) [22], [24], [37] are developed for large scale recognition tasks.

Most of current works in both categories still have some limitations. For example, the algorithms based on matching of constellation of local features can be applied to work with only small number of the objects in the database. Meanwhile, the recognition algorithms based Bag-Of-Visual-Words (BOVW) allow the large number of objects added into the database and it can recognize the object in an image very fast. However, one major limitation of these algorithms lies in the assumption that there must be only single object in the image. This requires some certain degree of user interactions e.g. specifying the region of interest.

In this research, we will comprehensively investigate the current state-of-the art algorithms for visual recognitions that are based on local appearance feature approach. The main goal of this research is to discover in detail the limitations of the current state-of-the art algorithms, and then to devise a novel algorithm for coping these shortcomings. The key aspects of our study include (but not limit to) Accuracy, Efficiency and Scalability. In addition to study the approach itself, we also investigate two different but related recognition problems including specific object recognition and object-category recognition. These are the problems that major current algorithms exploit the local feature based approach to tackle the problems.

คำนำ

เนื้อหาในรายงานวิจัยฉบับนี้ เป็นเนื้อหาที่นำมากจากบทความวิจัยเรื่อง Scalable Detection of Multiple Specific Objects using Bag-Of-Visual-Word Image Retrieval with Hough Voting based Ranking ที่ผู้วิจัยและรองศาสตราจารย์ ปกรณ์ แก้วตระกูลพงษ์ ได้จัดทำเพื่อให้พิจารณาส่งตีพิมพ์ในวารสาร Computer Vision and Image Understanding เมื่อวันที่ 6 กุมภาพันธ์ 2555 เนื้อหาของบทความกำลังอยู่ในช่วงพิจารณาเนื้อหาทางเทคนิคโดยทีมงานบรรณาธิการของวารสาร โดยเนื้อหาของบทความนี้เกิดอันเนื่องมากจากผลจากการทำวิจัยในโครงการนี้

ผู้วิจัยหวังเป็นอย่างยิ่งว่างานวิจัยนี้จะเป็นประโยชน์ต่อวงการวิจัยทางด้าน Computer vision เพื่อที่จะได้นำผลการวิจัยนี้ไปใช้ให้เป็นประโยชน์ต่อไป

ประดิษฐ์ มิตราปิยานุรักษ์

17 กุมภาพันธ์ 2555

ประโยชน์ต่างๆ อันเนื่องมาจากผลงานวิจัยในโครงการนี้

ผู้วิจัยขออุทิศให้

นางสาวเสาวลักษณ์ มิตราปิยานุรักษ์

Elsevier Editorial System(tm) for Computer Vision and Image Understanding
Manuscript Draft

Manuscript Number: CVIU-12-54

Title: Scalable Detection of Multiple Specific Objects using Bag-Of-Visual-Word Image Retrieval with Hough Voting based Ranking

Article Type: Regular Paper

Keywords: recognition, detection, image retrieval, scalable, large scale, bag of visual word, hough voting, a contrario

Corresponding Author: Dr. Pakorn Kaewtrakulpong, Ph.D.

Corresponding Author's Institution: King Mongkut's University of Technology Thonburi

First Author: Pradit Mittrapiyanuruk, Ph.D.

Order of Authors: Pradit Mittrapiyanuruk, Ph.D.; Pakorn Kaewtrakulpong, Ph.D.

Abstract: We present an algorithm for simultaneously recognizing and localizing planar textured objects in an image. The algorithm can scale efficiently with respect to a large number of objects added into the database. In contrast to the current state-of-the-art on large scale image search, our algorithm can accurately work with query images consisting of several specific objects and/or multiple instances of the same object.

Our proposed algorithm consists of two major steps. The first step is to generate a set of hypotheses that provide information about the identities and the locations of objects in the image. To serve this purpose, we extend Bag-Of-Visual-Word (BOVW) image retrieval by incorporating a novel re-ranking scheme based on hough voting technique. Subsequently, in the second step, we propose a verification algorithm based on RANSAC homography estimation in conjunction with a contrario based decision framework to draw out the final detection results from the generated hypotheses.

We demonstrate the performance of the algorithm on the scenario of recognizing CD covers with a database consisting of more than ten thousand images of different CD covers. Our algorithm yield to the detection results of more than 90% precision and recall within a few seconds of processing time per image.

Srinakharinwirot University

114 Sukhumvit 23,

Bangkok 10110, Thailand

February 6, 2012

Dear Professor Avinash C. Kak

Editor-in-Chief, CVIU

Please consider our submission for publication in Computer Vision and Image Understanding.

Some of our graphical plots include color for better discrimination between the curves. Should our paper be accepted and if the editors so desire, we will be glad to redo those plots in black and white.

Please note that this work has not been submitted for publication elsewhere.

Sincerely,

Pradit Mittrapiyanuruk

praditm@swu.ac.th

Srinakharinwirot University

***Highlights**

- Algorithm can accurately work with query images consisting of multiple objects
- Hypothesis generation based on Bag-Of-Visual-Word (BOVW) image retrieval
- Incorporate a novel re-ranking scheme based on hough voting technique
- Hypothese verification based on a contrario based decision framework

Scalable Detection of Multiple Specific Objects using Bag-Of-Visual-Word Image Retrieval with Hough Voting based Ranking

Pradit Mittrapiyanuruk[a], Pakorn Kaewtrakulpong[b,*]

[a]*Srinakharinwirot University, Thailand*
[b]*King Mongkut University of Technology Thounburi, Thailand*

Abstract

We present an algorithm for simultaneously recognizing and localizing planar textured objects in an image. The algorithm can scale efficiently with respect to a large number of objects added into the database. In contrast to the current state-of-the-art on large scale image search, our algorithm can accurately work with query images consisting of several specific objects and/or multiple instances of the same object.

Our proposed algorithm consists of two major steps. The first step is to generate a set of hypotheses that provide information about the identities and the locations of objects in the image. To serve this purpose, we extend Bag-Of-Visual-Word (BOVW) image retrieval by incorporating a novel re-ranking scheme based on hough voting technique. Subsequently, in the second step, we propose a verification algorithm based on RANSAC homography estimation in conjunction with *a contrario* based decision framework to draw out the final detection results from the generated hypotheses.

We demonstrate the performance of the algorithm on the scenario of recognizing CD covers with a database consisting of more than ten thousand images of different CD covers. Our algorithm yield to the detection results of more than 90% precision and recall within a few seconds of processing time per image.

Keywords:

*Corresponding author
Email addresses: praditm@swu.ac.th (Pradit Mittrapiyanuruk), pakorn.kae@kmutt.ac.th (Pakorn Kaewtrakulpong)

recognition, detection, image retrieval, scalable, large scale, bag of visual word, hough voting, a contrario

1. Introduction

The current research on object recognitions can be divided into two major groups [1]:(i) specific object (instance) recognition, and (ii) object-class recognition. Our proposed algorithm presented in this paper is along the line of recognizing specific objects. Particularly, we concern in developing a scalable recognition algorithm that can simultaneously identify the identities and localize the locations of multiple planar textured-rich objects in an image. The algorithm can scale efficiently with a growing number of objects added into the database. This kind of object recognition can be applied to various applications e.g. recognition of CD-covers [2], [3], recognition of book covers/book splines [4], [5], [6], detecting objects in household environments [7], [8], [9], and Augmented Reality [10].

Recently, some works [7], [11], [12], [8] based on matching of local features (e.g. SIFT and SURF) were proposed to simultaneously recognize and localize multiple objects in images. However, these works can only perform efficiently with a small number of the objects in the database.

On the other hands, some works based on Bag-Of-Visual-Word (BOVW) image retrieval [13], [14], [15] were developed for large scale recognition tasks. These algorithms allow a large number of objects added into database. And, they can efficiently recognize the object in a test (query) image. Nevertheless, one major limitation of these works lies in the assumption that there should be only a single object in the query image. Typically the accuracies of these works are dropped significantly when they are applied to images consisting multiple objects and cluttered backgrounds. This limitation was recently addressed in several works [3], [16], [17], [18], [9], [6].

The main reason that the current state of the art BOVW image retrieval performs inaccurately on query images containing multiple objects can be explained as follows. Basically these image retrieval engines will search for a single object in database that is most similar to the visual contents extracted from the whole query image (i.e., collection of visual words without considering their spatial information). As the query image consists of the visual contents mixing from multiple specific objects, this incidence usually leads the retrieval engine to ranking irrelevant database objects in high orders. That is, we tend to obtain the retrieval results in which the relevant

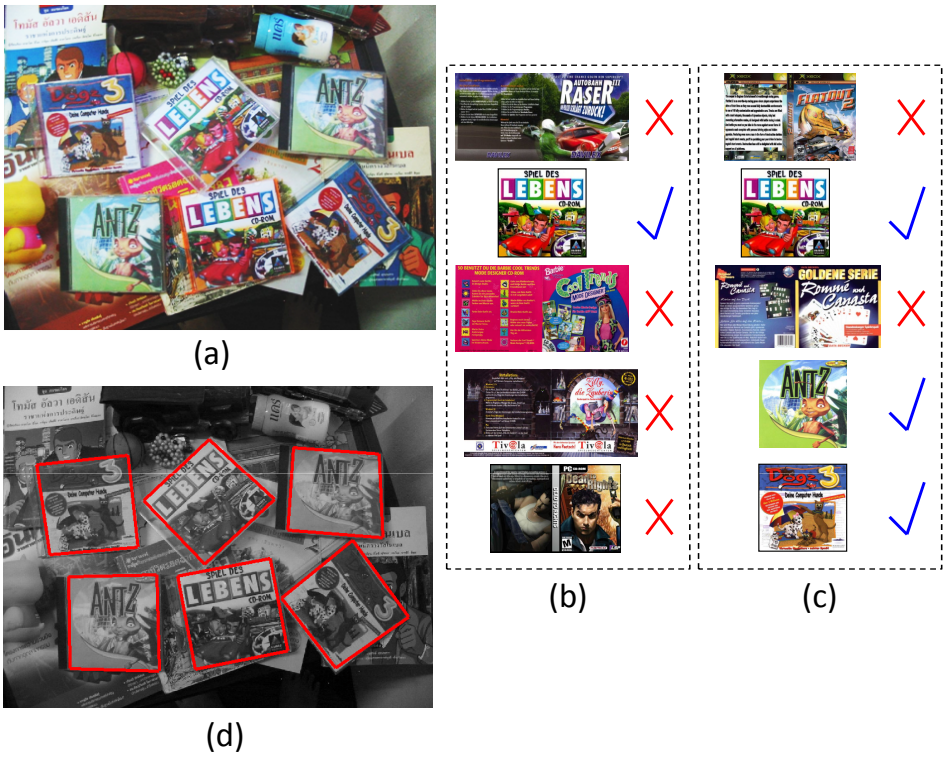


Figure 1: Issue in scalable recognition of multiple objects: (a) A query image with 3 different specific objects, each of 2 instances. (b) Top 5 retrieval result from TF-IDF image retrieval [14]. (c) Top 5 retrieval result from hypothesis generation of our method. (d) Localization result from hypothesis verification step. The bounding boxes of detected CD covers are shown in red.

database objects are ranked in very low orders of the recognition result (e.g. not in top 100 rankings). And, due to time processing constraint, these low ranked relevant objects will not usually be considered by a re-ranking module (e.g. geometric alignment based re-ranking [15]) in the post-processing step. This causes the retrieval engines to miss a large number of detections that finally yields to low recall rates.

An example of the aforementioned problem can be illustrated in Figure 1. In this figure, there are totally 6 objects that simultaneously appear in the query image shown in Figure 1(a). Specifically there are three different objects, i.e., CD-covers entitled *LEBENS*, *ANTZ*, and *Dogz 3*, where two instances for each specific object are located in different parts of the image. By applying a TF-IDF image retrieval [14], the top 5 most relevant database

objects are listed in Figure 1(b). We found that only one object (*LEBENS* CD) is listed in the second rank. Meanwhile the other two objects are ranked in the low orders, i.e. the 45^{th} and the 407^{th} orders (out of 11,444 database images), respectively.

To tackle the aforementioned limitation, we propose an algorithm that is extended from the state-of-the-art Bag-Of-Visual-Word (BOVW) image retrieval by incorporating a novel ranking scheme based on hough voting technique. Regarding to the same query image shown in Figure 1(a), our proposed ranking scheme can efficiently retrieve all relevant objects in the top 5 ranking orders (2^{th}, 4^{th} and 5^{th}) as shown in Figure 1(c). Furthermore, only the relevant objects were detected along with their locations after applying the verification step of our algorithm as shown in Figure 1(d).

There exists some works [4], [16], [3], [19], [10], [20] that stay along the same line of our work presented in this paper. The brief reviews of these relevant works will be presented in the next section. Regarding to the literature, the contribution of our work can be summarized as follows.

- We propose a novel ranking scheme based on Hough voting for large scale Bag-Of-Visual-Word image retrieval [13], [14], [15], [21]. This part is used for the object hypothesis generation step of our algorithm where the hypotheses are derived from top rank image retrieval results. With respect to the current state-of-the-art, our algorithm can accurately draw a candidate list of relevant database objects in test images that consists of multiple objects (different specific objects and/or multiple instances of the same objects) in the presence of clutter background.

- We propose a hypothesis verification based on *a-contrario* decision framework [22] [23] [24]. Specifically, we propose to use *Number of False Alarms (NFA)* as the hypothesis quality score in our proposed greedy based hypothesis selection. This verification step is applied with the top rank candidates of generated hypotheses to make a final decision on the identities and the locations of objects in the image.

The remainders of this paper are organized as follows. We present some related works in Section 2. The overview of our proposed algorithm is explained in Section 3. The details of the algorithm are presented in Section 4 and 5. Finally, the experimental result is reported in Section 6.

2. Related Work

In [7], the author proposed an object recognition algorithm based on matching of local features, i.e., SIFT. The algorithm can recognize multiple objects simultaneously appeared in images. In this work, an object is represented with a set of SIFT features extracted in a model image. To recognize the object in a test image, first the matchings between the model SIFT features and the SIFT features extracted in the test image are established. Then a geometric verification based on Hough transform and RANSAC is applied to identify the object identities and their corresponding poses in the image. In [25], the authors proposed to apply an unsupervised clustering on hough space generated from SIFT matching. However, this work can be applied to detect only one specific object in images. In the similar spirit to [7], some works [11], [12], [8] were also proposed. Generally, the main drawback of the approach based on matching of local features is that the algorithm could run very slowly with respect to a growing number of model objects added into the database.

In [19], the authors addressed the issue of scalability by representing local features with codewords obtained from applying a vector quantization to the set of local features. However, they showed the experiment with only 50 objects in the database. Conceptually, this work can not work efficiently with a large scale database, e.g., more than 10,000 model objects.

To tackle the issue in large scale object recognition, the authors in [13] proposed to use both the notion of vector quantization on local features and the idea of TF-IDF indexing on quantized features (which is borrowed from text retrieval area) to efficiently search for near-duplicate frames in videos. This approach of image retrieval is usually referred to as Bag-Of-Visual-Words (BOVW) approach. This idea was strengthen in [14] with a faster vector quantization scheme referred to as Vocabulary Tree. Relatively, in [15] the authors proposed to use Approximate K-Means clustering for feature quantization.

As mentioned in the previous section, these BOVW retrieval techniques work inaccurately in the case of query images containing multiple objects and clutter background. Several modifications were proposed to tackle this problem. The most relevant works to ours are the ones proposed in [16], [3], [9], [6], [4].

In [16], the authors resorted to a large scale object image retrieval using Vocabulary Tree [14]. However, they proposed to modify the ranking score

by incorporating a weight obtained during feature quantization step. To our best knowledge, this scheme did not directly address the issues of multiple objects and clutter background in images. Furthermore, they exploited a sliding window based approach to localize the bounding boxes of the objects in the image. Generally, one disadvantage of sliding window approach is that the algorithm could run very slow even though it is applied as a post-processing step.

In [3], the authors resorted to a variation of weak geometric consistency [21] in ranking relevant object images. They used a 2D vote space on key-point orientation and scale in the ranking process to retrieve a set of top most relevant images. Our proposed ranking scheme in this paper is very similar to this relevant work. However, we propose to adopt a voting scheme presented in [26], [27] in conjunction with Inverse-Document Frequency (IDF) in the process of ranking the topmost relevant database object images. In the case of query images with multiple objects, we found that our scoring yield to a better recall than the ranking that only uses the number of votes corresponding to the peaks of voted spaces as proposed in [3].

In [18], [9], [6], some variations of unsupervised clustering techniques are applied to the features extracted in query images. Then a TF-IDF image retrieval is applied to each individual found cluster. The recalls of these approaches are very dependent on how robust the feature clustering steps could perform. Unlike these works in which the clustering performs in an unsupervised manner, our work presented here exploits the knowledge of objects in feature grouping in which the contributions of features to possible object candidates are dependent on the objects that these features are voted for (see more explanation in next sections).

In [4], the authors proposed an algorithm based on data-dependent multi-class branch-and-bound framework for multiple object recognition/localization. The authors performed the experiments on the database of different 100 book images. However, the drawback of this work lies in the scalability issues in which it fails to work efficiently with a large size of database e.g. 10,000 images.

Regardless of scalability issue, Rabin et al [28] proposed an algorithm for recognizing multiple objects by extending the notion of *a contrario* RANSAC [29], [30], [31]. Our work shares the same idea in which we exploit the quantity *NFA* as the hypothesis quality score in our hypothesis verification step.

Some relevant works on detection of multiple object classes (i.e., object

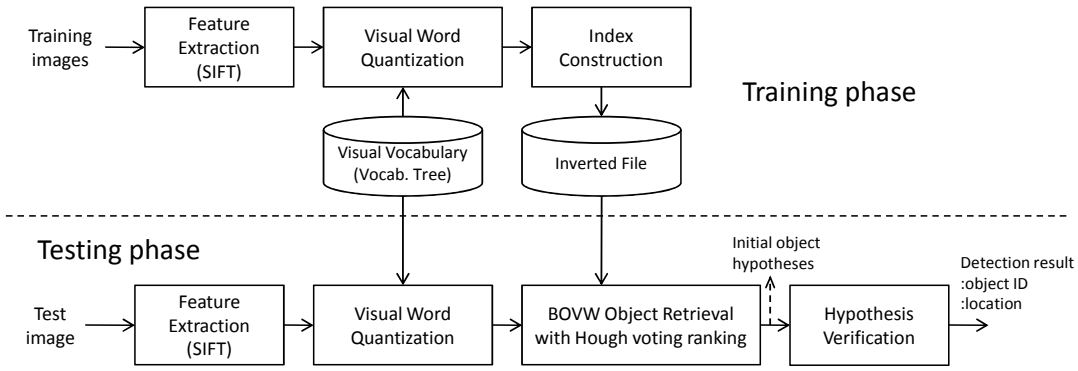


Figure 2: Algorithm Overview

category detection) are proposed in the current literature. Leibe et al [32] proposed the state-of-the-art detection framework, referred to as Implicit Shape Model (ISM). This method is an extension of Generalized Hough Transform (GHT) in which the results of local feature matching are used to vote for the object position in the image. In [33], [34], the methods extended from ISM are proposed to tackle the problem of multiple object class detection. Recently, some work [35] based on sliding window approach is also proposed.

3. Overview of the algorithm

The block diagram of our algorithm is depicted in Figure 2. An overview to the our algorithm can be explained as follows:

1. *Training phase*: the training images of all target objects are enrolled into the system. Associated with each image, we also assume that a bounding box of the object is annotated. Then, the SIFT features are extracted from the images. Each feature is quantized to represent as a visual word by using Vocabulary Tree [14]. Then an object is modeled with a set of keypoints derived from SIFT. Each keypoint is specified with 5 entries: x, y, scale, orientation and visual word ID. Finally, an inverted file based indexing structure [13] is constructed from the keypoints of all training object images.
2. *Testing phase*: given a query (test) image which may contain multiple relevant objects and clutter background, we want to determine the

identities and the 2D locations of objects using the model constructed in the training phase. This is achieved by the following steps.

(a) Feature Extraction: A set of SIFT keypoints are extracted from the test image. The visual words are obtained by quantizing the SIFT descriptors of keypoints.

(b) Hypothesis generation: A set of initial object hypotheses are obtained by using BOVW image retrieval with Hough voting ranking. Each hypothesis is specified with object ID and image location. To fulfill this step, we first establish a set of matchings between the test image keypoints and database object (training images) keypoints in which a matching is declared if the keypoints belong to the same visual word. These matchings are efficiently retrieved with the help of the inverted file indexing structure. Then each match will cast a vote for hypothesis in a 3D voting space implemented in the form of accumulator bins (2D for image location and 1D for object ID). We use IDF weight [13] of each visual word as the vote weight to collect in these bins. Finally a set of candidate hypotheses are generated from the object identity and the keypoint matchings that are corresponding to the peaks in the voted bins whose scores are ranked in the topmost orders (e.g. 50 highest peaks).

(c) Hypothesis verification: a RANSAC based estimation of planar transformation (e.g. homography, affine) in conjunction with an *a-contrario* decision framework [30] is applied to draw out detection results from the generated hypotheses. Specifically, the algorithm is based on the key idea that a test image keypoint could contribute to only one single object. Then, an object instance is selected one by one from a set of ranked hypotheses in a greedy iterative manner where the quantity, referred to as *NFA* (derived from *a-contrario* decision framework [30]), is used as hypothesis quality score.

4. Hypothesis Generation

In this section, we will explain in details of our proposed Bag-Of-Visual-Words (BOVW) object retrieval with Hough voting ranking that works as the hypothesis generation step of the overall algorithm. It will generate a set of hypotheses that are the candidates for the detected instances of objects

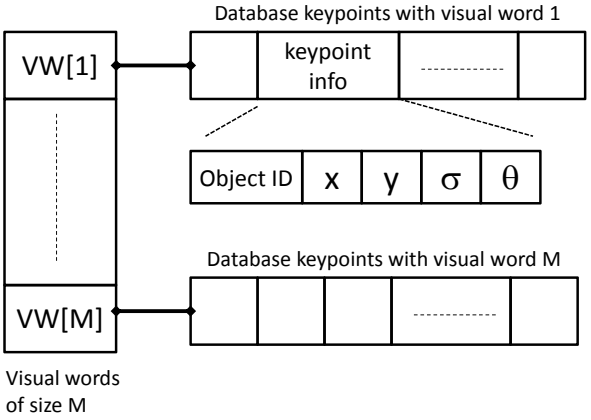


Figure 3: Inverted file based index structure

in the test image. Each hypothesis is defined with an object identity and its image location. Subsequently, the hypothesis verification step explained in the next section is applied to each hypothesis to make a final decision of whether a relevant database object really exists.

As mentioned previously, most state-of-the-art algorithms on BOVW image retrieval [13], [14], [15], [21] resort to Term Frequency-Inverse Document Frequency (TF-IDF) to rank relevant object images. This ranking scheme is usually failed in the case of multiple objects and clutter background in query images [3], [9], [6]. In this paper, we tackle the aforementioned shortcoming by ranking relevant images based on scores calculated in a hough voted space. Our proposed idea for this part of the algorithm is in the same spirit of weak geometric consistency as proposed in [21], [36], [37], [3]. The details of the algorithm for this part consist of the following steps.

4.1. Match

For each keypoint extracted in the test image, we retrieve a set of matches to the keypoints of relevant objects in the database (i.e., training images). These matches are efficiently drawn with the help of the inverted file based indexing structure [13]. This index is constructed from the keypoints of all training images in the training phase. As illustrated in Figure 3, the inverted file index is abstractly a table of size equal to the number of visual words in the vocabulary used in the quantization step. The data filled in each row entry of table is a collection of information about training image keypoints that belong to the corresponding visual word. Specifically, the information

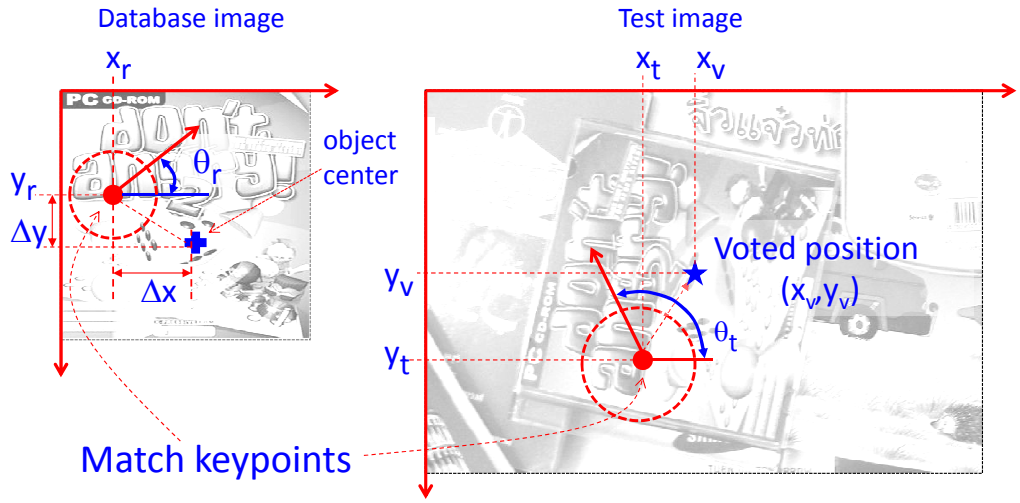


Figure 4: Illustration of voted location

kept in each entry consist of an integer for object identity and four floating-points numbers for keypoint location (x,y), scale (σ), orientation (θ) that appeared in the training image. The values of x, y, σ, θ will be used in the hough voting based scoring explained next. It is noteworthy to mention that our algorithm does not use the SIFT descriptors of keypoints in the matching process. Specifically we do not store the SIFT descriptors of training image keypoints into the database.

4.2. Vote

Each match found in the previous step will cast a vote for a candidate hypothesis in 3D hough space corresponding to object identity (O_v) and object center location (x_v, y_v) in the test image. This voted space is represented in the form of 3D accumulator bins. Similar to [34], we can consider this 3D voted space as multiple separated 2D vote spaces where each 2D voted space is corresponding to the voted space for object center location of a particular object identity.

For estimated object center locations, we resort to the (in-plane) rotation invariant voting scheme as suggested in [26], [27] (Note that some variations of the voting scheme are also proposed in [38], [39]). That is, see Figure 4 for an illustration, a match between the test image keypoint specified with (x_t, y_t, σ_t, θ_t) and the database keypoint (x_r, y_r, σ_r, θ_r) of the object identity O_v will cast a vote for the entry (O_v, x_v, y_v) of 3D voted accumulator. The voted

object center location (x_v, y_v) can be calculated by the following equations.

$$x_v = x_t + \frac{\sigma_t}{\sigma_r} * (\sqrt{\Delta x^2 + \Delta y^2}) * \cos(\theta + \theta_r - \theta_t) \quad (1)$$

$$y_v = y_t - \frac{\sigma_t}{\sigma_r} * (\sqrt{\Delta x^2 + \Delta y^2}) * \sin(\theta + \theta_r - \theta_t) \quad (2)$$

where $\theta = \tan^{-1}\left(\frac{\Delta y}{\Delta x}\right)$. And, Δx and Δy are the x- and y-distances between the database keypoint location to an arbitrary object reference location in the training image.

In Figure 5, we show some graphical examples of the voted locations (star markers) that are generated from the correct matches in Figure 5(a) and the voted locations obtained from incorrect matches in Figure 5(b). Typically, the voted locations from correct matches tend to cluster at some position (see the grouping of star markers in 5(a)).

In our implementation, we use the object center location derived from the centroid of bounding box vertices that are annotated with the training image. Furthermore, the values of Δx, Δy and θ can be pre-computed and be kept in the corresponding entry of inverted-file based index during the training phase. However, now we decide to recompute these quantities for the purpose of memory space saving.

To further reduce memory usage requirement, we quantize the location (x_v, y_v) to vote for a coarser resolution accumulator bin. That is, we make a single voted location bin to cover 20x20 pixels. Therefore, for the image of size 640x480 pixels, this will require a 2D array of size 32x24 for the accumulator of one object identity. In the same spirit as [7], we diminish the boundary effects in bin quantization by additionally voting each match to neighborhood bins (± 1) in voted x and y dimensions.

Moreover, we also take into account for the effect of inter-image burstiness [40] in which some visual words may appear more frequently than others across database images. This is achieved by incorporating a vote weight obtained from IDF weight of the corresponding visual word. That is, a match of visual word v_i will cast a vote for the location defined in (1) and (2) with the weight w_i defined by:

$$w_i = log(\frac{N}{N_i}) \quad (3)$$

where N is the total number of the objects in the database, and N_i is the number of the database objects that consists of at least one visual word v_i.

4.3. Find peaks

Typically if there exists an instance of database object in the test image, the weights accumulated in the voted space of the corresponding object identity will form some peak in the vicinity of estimated object center location. This peak will correspond to a hypothesis. An illustrated examples of voted spaces generated from the matches in the image of Figure 1 is shown in Figure 6. In these pictures, we show 5 different voted spaces of top 5 objects retrieval results. Particularly, we visualize the voted weights as pixel intensities (0-255) in image space where the areas with lower intensity (darker) indicate stronger weight values than brighter areas. From the pictures, it is noteworthy to mention that the peaks for voted spaces in Figure 6-b and 6-d are not corresponding to any actual object in the image. Subsequently, these incorrect hypotheses corresponding to the peaks will be rejected by the verification step as explained in the next section.

To finish the hypothesis generation task, the algorithm in this step will identify the potential peaks in the voted spaces. In this work, we apply a conventional Non-Maximum Suppression (NMS) to select the topmost peaks. Particularly, we opt to draw out the hypotheses corresponding to the topmost 50 peaks (i.e., we assume that less than 50 object instances simultaneously appear in an image). However, this number can be adjusted with some slightly additional costs of computation (e.g. changing from 50 to 100 could increase less than two times in processing time for the verification step).

For each identified peak, we generate a candidate hypothesis that is defined by an object identity and an estimated object center location corresponding to the peak location. We also accompany the estimated object center location with the supporting keypoint matches that vote for the peak. Explained in the next section, these keypoint matches of each candidate hypotheses will be used by a geometrical alignment based verification algorithm to locate an exact location of object in the image.

5. Hypothesis Verification

From the previous step, we obtain a set of hypotheses where each hypothesis is specified with an object identity and a set of putative keypoint matches that is voted for the corresponding peak in voted spaces. Mathematically, we use the following notations to represent the hypotheses.

- $H = \{h_k; k = 1, ..., L\}$ be the set of generated hypotheses where L is the number of hypotheses.

- $h_k = \{(o_k, P_k)\}$ be the k^{th} hypothesis where o_k is the object identity and P_k is the set of putative keypoint matches.

- $P_k = \{(p_{k,i}, p'_{k,i}); i = 1, .., M_k\}$ be the set of putative matches between the database (training image) keypoints and the test image keypoints where M_k is the number of keypoint matches.

- $p_{k,i}$ and $p'_{k,i}$ are the database keypoint and the test image keypoint of match i in hypothesis k, respectively.

In this stage, we will make a decision whether the likelihood of each generated hypothesis is strong enough to declare the presence of an object instance. This step will also resolve the ambiguity from conflicting hypotheses [32], [33].

We consider that any two or more candidate hypotheses are conflicted to each other if these hypotheses share some test image keypoints in their putative matches. Generally the bounding boxes in the test image of these conflicting hypotheses will overlap to each other that lead to two possible cases. The first case is that one of conflicting hypotheses is corresponding to an actual object instance while the remaining hypotheses are due to false positives. The second possibility is that all of these conflicting hypotheses are corresponding to false positives and should be rejected.

An example of conflicting hypotheses in a test image can be shown in Figure 7. In this image, it consists of only one object instance. However, we found three conflicting hypotheses that partially share putative keypoint matches. The first hypothesis (Spiderman CD-cover) is corresponding to the actual object instance, i.e., true positive. Meanwhile, the second and third hypotheses are corresponding to false positives. These three hypotheses share the test image keypoints in the vicinity of *"PlayStation"* logo of CD-cover. The corresponding bounding boxes of these hypotheses are also overlapped to each other. These bounding boxes are visualized with three different colors i.e., red, green and yellow in which the red bounding box is the correct one.

The key idea of our proposed verification step is that we measure the likelihood of a hypothesis based on its inlier matches. These inlier matches are obtained by applying a RANSAC based estimation of planar transformation [41] to the putative matches of hypothesis. To rate the qualities of hypotheses, we resort to the *a-contrario* based decision framework [22] [23] [24] in which we use the quantity, referred to as *Number of False Alarms (NFA)*, computed from the inlier matches as hypothesis quality scores. Finally, to

resolve conflicting hypotheses, we propose a greedy based iterative scheme for selecting plausible hypotheses.

5.1. Hypothesis quality score based on A-Contrario decision framework

In short, the *a-contrario* decision methodology [22] [42] [23] [24] is based on the Helmholtz principle in which a geometrical based event in an image will be perceived if the likelihood that the corresponding event occurs by chance is very low. Specifically, the methodology associates a computational quantity, referred to as *Number of False Alarms (NFA)*, to the geometric event. The *NFA* of an event is defined as the expectation of number of occurrences of the event under a background (noise) model. This background model is referred to as *a-contrario* model. In general, an event with very low value of *NFA* is considered as meaningful event.

In the literature, there are several variations on works based on *a-contrario* model in which basically they are different on how to mathematically define and compute *NFA* that are exploited in various application contexts e.g. feature grouping [22], [43], motion and change detection [44], [45], feature matching [46], [47], etc. In our algorithm, we resort to the idea of using *a-contrario* model for estimating geometric transformations between two images as suggested in [29], [28].

That is, we define the *NFA* based on the set of inlier matches between the database keypoints and the test image keypoints associated with a hypothesis. This set of inlier matches is obtained after applying a RANSAC based estimation to the putative matches corresponding to the hypothesis. To show the mathematical formulae for computing *NFA*, we will use the following notations.

- $S = \{(m_i, m_i') | i = 1,\ 2,\ .., N\}$ be the set of inlier matches of keypoints of a hypothesis where m_i and m_i' are the database keypoint and the test image keypoint, respectively.
- $S' \subseteq S$ be the minimal sample subset (MSS) that is used to compute the 2D planar transformation by RANSAC based estimation.
- $c = |S'|$ where $c = 3$ for affine and $c = 4$ for homography.
- $\mathbf{A}$ be the 2D planar transformation computed from S'.
- $S \backslash S'$ be the set of inlier matches that exclude S'

- k be the cardinality of the set $S \backslash S'$, i.e., $k = |S \backslash S'|$.
- $\alpha(S, \mathbf{A}, S')$ is the rigidity of S which is defined as

$$\alpha(S, \mathbf{A}, S') = \max_{(m,m') \in S \backslash S'} \left\{ \frac{\pi}{w'h'} d(\mathbf{A}m, m')^2 \right\} \tag{4}$$

  where w' and h' are the width and height of the test image.
- $d(\mathbf{A}m, m')$ is the 2D distance between the test image keypoint m' and the back-projection of corresponding database keypoint m by the transformation $\mathbf{A}$.

By using the above notations, the number of false alarm (*NFA*) of a hypothesis according to S, $\mathbf{A}$ and S' can be defined as

$$NFA(S, \mathbf{A}, S') = (N - c) \begin{pmatrix} N \\ k \end{pmatrix} \begin{pmatrix} k \\ c \end{pmatrix} [\alpha(S, \mathbf{A}, S')]^{k-c} \tag{5}$$

Given a threshold ε, if we can find the match (m, m') that maximizes $\alpha(S, \mathbf{A}, S')$ in (4) and also makes $NFA(S, \mathbf{A}, S') \leq \varepsilon$; consequently we consider the set of matches (i.e., the hypothesis) to be ε-meaningful. This implies that this set of inlier matches is likely to be the matches corresponding to an actual object instance in the test image. Moreover, the smaller that the value of *NFA* is will indicate that the corresponding hypothesis is more meaningful. In general, the value of this threshold set (ε) can be fixed to 1. Therefore, if the *NFA* value computed from the matches of a hypothesis is larger than 1, we reject the hypothesis.

In our algorithm, we intent to make a larger value of hypothesis quality score means that the hypothesis is stronger (i.e., more likely to be correct). Therefore, we define the quality score of a hypothesis by the negative of logarithm of NFA. That is, for the NFA of a given hypothesis h_k, the quality score is given by:

$$Q(h_k) = -\log(NFA) \tag{6}$$

Note that, the main reason of computing the logarithm of *NFA* is that computing the quantity in (5) usually involves with the computation of very small numbers (e.g. $< 10^{-30}$). Due to numerical accuracy issue, most *a-contrario* approaches look forward to compute the value of $\log(NFA)$ instead of directly computing NFA in their implementations.

5.2. Verification algorithm

As mentioned earlier, a test image keypoint may match to multiple keypoints of several database objects during the matching step of Hypothesis Generation (Section 4). From this effect, it is likely that the set of putative matches associated to two or more hypotheses may partially share the set of keypoints in the test image. This leads to the ambiguity on conflicting hypotheses that need to be resolved appropriately.

To tackle this issue, we propose a greedy based algorithm that iteratively selects the potential hypotheses and prunes out the conflict ones. Before going into details, we want to emphasize that our algorithm is based on the key assumptions as follows. First we assume a keypoint in the test image can contribute to only a single object instance. Second, we also assume that there is no partial occlusion in the test image. Although we found that our algorithm can resist to some certain degree of partial occlusions, we leave this problem as a future work.

Expressed with a psuedo-code in Algorithm 1, our verification algorithm performs greedily in the hypothesis selection in which a single best hypothesis at each iteration is pulled out from the list of candidate hypotheses. To compare among hypotheses, we adopt the *NFA* based quality score as explained in the previous section. If the quality score of selected hypothesis is still greater than a fixed threshold, we add the hypothesis into the detection result. Before starting the next iteration, we remove the matches, that are associated with the test image keypoints of the selected hypothesis from other remaining hypotheses. We iterate these steps until no hypothesis whose score is greater than the threshold is found.

An example of the iterations of verification algorithm on the test image in Figure 1 can be illustrated in Figure 8. In each image, it showed the detected instance of an object that the verification algorithm returned at each iteration in blue bounding boxes. The keypoints extracted in images are shown with the yellow markers. We also showed the matches (red lines) between the test image keypoints and the keypoints of database image on the left hand side. Note that the incorrect hypotheses corresponding to Figure 8(b) and 8(d) are rejected by the verification algorithm.

The details of the algorithm can be explained as follows. We start with all hypotheses to be in the list of active hypotheses H. For each of active hypothesis i in H, we determine the set of inlier matches S_i from its putative matches P_i. This is accomplished by the procedure *RansacPlanarTrans-Estimation* (Line 7) in which a RANSAC based estimation of 2D planar

transformation (i.e., affine or homography) is carried out. This procedure will return three outputs, i.e., the set of inlier matches S_i, the estimated transformation $\mathbf{A}_i$ and the set of sample matches S_i' that is used to compute $\mathbf{A}_i$. In addition, we also incorporate a rigidity constraint as suggested in [48] into our RANSAC homography estimation. The main purpose of enforcing the constraint is to prevent non-physically meaningful homographies. From the three outputs returned by the RANSAC based homography estimation, we compute the *NFA* of the hypothesis according to (5) by the procedure *ComputeNFA* (Line 8) and then compute the hypothesis quality score $Q(h_i)$.

Next, we find the best hypothesis whose quality score is maximum among all active hypotheses in H. This is corresponding to Line 13 of the pseudo-code. If the quality score of the best hypothesis is greater than the value of $-\log(\epsilon)$, this means that $NFA < \epsilon$. Therefore we can conclude that the hypothesis is meaningful and can be accepted. The bounding box of the detected object instance in the test image can be determined by the back-projection of database object bounding box, denoted by *DbObjBB*, into the test image coordinates by using the estimated transformation $\mathbf{A}_j$, as in Line 21. Then the projected bounding box BB_j and the corresponding object identity o_j will be added into the detection result L, as in Line 22. With regard to the example in Figure 8, this step can be seen in Figure 8(a). The first object instance (*LEBENS* CD-cover) is detected where the bounding box is shown in blue color.

After that, the accepted hypothesis will be removed from the set of active hypotheses, as in Line 23. At the same time, we will remove the putative matches of other hypotheses that the supporting test image keypoints fall into the bounding box BB_j of the accepted hypothesis. This step is to cover the assumption we make earlier that a single keypoint in the test image can contribute to only one object instance. Regarding to the pseudo-code, this step is corresponding to Lines 24 to 26 in which any match of other remaining hypotheses that falls into this projected bounding box will be removed from the hypotheses by the procedure *RemoveMatch*.

Then, the algorithm will start the next iteration with the updated information to detect other object instances. This step can be seen in Figure 8(b). The matches within the bounding box of the first detected object instance (8(a)) are removed. Then the algorithm continues and can detect the subsequent object instance (*ANTZ* CD-cover).

Subsequently, the algorithm will continue the iteration process to detect other objects as shown in Figures 8(b) to 8(d). Whenever there is no any

hypothesis whose quality score is larger than the threshold, we have no any good hypothesis that can be accepted as an object instance. Thus we will stop the iteration by setting the flag *stop* to be *true* in Line 28. The final result in L is the list of object instances detected in the test image.

6. Experiments

6.1. Implementation details

We implemented the proposed algorithm by using several toolkits. For feature extraction, we use *VLFeat* [49] to extract SIFT features in images. We exploit *VOCSEARCH* [50] for the purpose of visual word quantization with vocabulary tree [14] and inverted file indexing [13]. We also used the pre-trained vocabulary tree consisting of 1M visual words that is provided by *VOCSEARCH*. We implement BOVW object retrieval with Hough voting (Section 4) in C++ by modifying the source code provided by the *VOCSEARCH*. Finally, our hypothesis verification (Section 5) is implemented with MATLAB in which we follow the guideline provided by Moisan et al in [31].

6.2. Datasets

We evaluate the proposed algorithm with the Caltech Game CD/DVD covers dataset [51]. This dataset consists of 11,400 images for CD/DVD covers of video games. We use these images as the training images. One training image is assigned to an unique object identity. The sizes of these images are about 400x400 pixels. We did not directly annotate the bounding boxes of the objects in these images. Instead, the vertices of bounding box of each object in an image is derived from the image boundary, i.e., $(0,0),(W-1,0),(W-1,H-1),(0,H-1)$ where W and H are the image width and height, respectively. After SIFT extraction, we found that roughly there are about 1,000 SIFT keypoints extracted in each training image. Some examples of training images can seen in Figure 9.

To collect the test images, we selected 50 selected CD images, printed out these images with a color laser printer in the actual size and put them into jewel CD cases. Then we took the pictures of these CD covers with a digital camera. The original resolution of images we capture is 1600x1200 pixels. However, for the purpose of processing time, we resize the images to 1024x768 pixels before applying the SIFT extraction module. The number of SIFT keypoints in each test image is in the range of 3,000 to 5,000 keypoints.

Table 1: Description of different test sets.

Test set	Number of different CD cover types in an image	Total number of CD covers in an image	Number of images
1	1	1	100
2	1	2	110
3	1	4	100
4	2	2	50
5	4	4	30
6	2	4	52
7	6	6	18
8	3	6	35

In other words, we can consider the selected 50 CD covers as the probe set and the remaining ones are the distracter.

We created 8 different sets for test images. The images in these sets consisted of a varying number of CD covers in images as explained in Table 1. In the test sets 1, 2 and 3, there are one, two and four of a specific type of CD cover in the test images, respectively. These test sets demonstrate the case of multiple instances of the same specific object in an image. For the test sets 4, 5 and 7, there are two, four and six CD covers of different types in each image. These test sets demonstrate the case of multiple specific objects in an image. For the test set 6, there are two instances of two different types of CD covers in an image (i.e., there are totally 4 CD covers in an image). For the test set 8, there are two instances of three different types of CD covers in an image (i.e., there are totally 6 CD covers in an image). The test sets 6 and 8 demonstrate the case of multiple instances/multiple specific objects in an image. Some examples of these images are shown in Figures 11, 12, 13 and 14. Totally, there are 495 images with 1,446 instances of CD-covers in all test sets. For the ground truth, we manually annotate an object identity and bounding box vertices for each CD cover instance in the test images. Note that all images are captured in color. However, we convert to gray scale images before applying our algorithm.

Table 2: Detection performance of our algorithm on different test sets

Test set	TP	FP	nP	$Precision$	$Recall$	AP
1	99	0	100	1.00	0.99	0.909
2	199	5	220	0.98	0.91	0.908
3	391	0	400	1.00	0.98	0.909
4	100	1	100	0.99	1.00	1.000
5	115	0	120	1.00	0.96	0.909
6	193	1	208	0.99	0.93	0.909
7	103	0	108	1.00	0.95	0.909
8	179	0	210	1.00	0.85	0.818
All (1-8)	1379	7	1466	0.995	0.941	0.909

6.3. Results

To evaluate the detection results, we follow the PASCAL Visual Object Class (VOC) challenge [52]. That is, we justify each detection result as either true positive or false positive. A detection result is considered as a true positive if (i) it has the same object identity as the one of ground truth, and (ii) ratio of intersection over union ov between the detected bounding box BB_d and the ground truth bounding box BB_{gt}, as expressed in (7), is larger than 0.5.

$$ov = \frac{area(BB_d \cap BB_{gt})}{area(BB_d \cup BB_{gt})} \tag{7}$$

We measure the accuracy of our proposed method on each test set with the precision and the recall as defined in the following equations.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{nP} \tag{9}$$

where TP is number of true positives, FP is number of false positives, nP is total number of positives in a test set.

The detection results of our algorithms on eight test sets are listed in Table 2. To plot a precision-recall (PR) curve, we rank the detection results

in descending orders according to the values of hypothesis quality scores (i.e., $-\log(NFA)$). The Precision-Recall curve for each test set is shown in Figure 10. In the same spirit to [52], we also computed the interpolated average precision (AP) and showed in the last column of Table 2. The AP is defined as the average of precision at eleven equally spaced recall level $[0, 0.1, 0.2, ..., 0.9, 1]$, which can be expressed as:

$$AP = \frac{1}{11} \sum_{r \in [0,0.1,...,1]} p_{interp}(r) \tag{10}$$

where $p_{interp}(r)$ is the interpolated precision at the recall level r which is defined by

$$p_{interp}(r) = \max_{\tilde{r}:\tilde{r} \geq r} p(\tilde{r}) \tag{11}$$

where $p(\tilde{r})$ is the measured precision at recall $\tilde{r}$.

Of all test sets, our algorithm can detect 1,379 instances of CD covers (number of true positives) from totally 1,466 CD-covers (number of positives) in all test images. Also, the algorithm returned very small number of false positives (7 out of 1,386 detections). Roughly, our proposed algorithm yields to more than 90% of recalls and precisions in all test sets. Besides by the main role of *a-contrario* based geometric verification, we found that by enforcing the rigidity constraint in RANSAC process as suggested in [48] to reject inconsistent homographies is also helpful for reducing number of false positives. This issue was also reported in some literature [31], [53].

Some of qualitative results are shown in Figure 11 (for the test sets 1 and 2), Figure 12 (for the test sets 3 and 4), Figure 13 (for the test sets 5 and 6) and Figure 14 (for the test sets 7 and 8). For the sake of clarity, we do not show the identities of objects. And, we display the projected bounding boxes of detected CD-covers in red.

6.4. Processing time

The processing time per image of our algorithm is presented in Table 3. Again, the sizes of test images is 1024x768 in which there are about 3,000 to 5,000 SIFT keypoints in an image. We measured the running times of algorithm on a PC with Intel Core 2 Duo 3GHz and 2GB of RAM. We want to emphasize that the verification step presented in Section 5 is still implemented with MATLAB. The actual processing time of this step can be improved with C++ implementation.

Table 3: Processing time of our algorithm

Step	Time per image (seconds)
SIFT extraction	1.32
Visual word quantization	0.26
Hypothesis generation	1.44
Hypothesis verification (in MATLAB)	1.13

6.5. Comparison with the state-of-the-art large scale image search

In this section, we report the comparative result between our proposed ranking scheme based on hough voting technique and the state-of-the-art large scale image search using TF-IDF ranking scheme with Vocabulary Tree that is proposed by Nister and Stewenius [14]. Note that our proposed ranking based on hough voting is used as the hypothesis generation step explained in Section 4. An illustrated example of the comparison is previously mentioned in Figure 1.

To compare these two approaches, we will evaluate how well the algorithms can retrieve a set of relevant database objects in the top ranked orders by measuring the recall (in percentage) of the results within the top database candidates. Specifically, in this comparison, we will not consider any object localization result that our algorithm also can produce after applying the verification step. That is a retrieval result will inform the candidate list of object identities in the query image. If the query image consists of multiple instances of a single specific object, we will consider them as one sample instance in the recall computation. Furthermore, we vary the number of top database candidates, i.e., N_{top} to be 20, 40, 60, 80, and 100. Thus, at any value of N_{top}, if the retrieval results in the top N_{top} ranked list hit the labels in the ground-truths, we will count these results in the recall calculation.

The results for this comparative evaluation on the data-sets in Section 6.2 can be shown in Table 4. We also illustrate the comparative result with the graphs in Figure 15. From the result, we found that our proposed ranking scheme outperformed the one proposed by Nister & Stewenius [14]. Our algorithm yield to above 90% of recall at all values of number of top database candidates. Meanwhile, in the test sets 5, 7 and 8, the Nister & Stewenius approach is completely failed to recognize the objects in which the recalls are less than 50% at all values of number of top candidates.

Table 4: Comparison of recall rates (in percentage) between our hough voting based ranking and the Nister & Stewenius (N-S) approach [14] at N_{top}=20, 40, 60, 80 and 100. (See also the graphs in Figure 15 for illustration.)

Test set	Ours					N-S approach				
	N_{top}					N_{top}				
	20	40	60	80	100	20	40	60	80	100
1	99.0	99.0	99.0	99.0	99.0	70.0	76.0	79.0	79.0	79.0
2	92.7	94.6	96.4	97.3	97.3	79.1	80.0	80.0	82.0	83.9
3	98.0	99.0	99.0	99.0	99.0	97.0	98.0	99.0	99.0	99.0
4	97.0	100.0	100.0	100.0	100.0	26.0	33.0	38.0	41.0	44.0
5	93.3	95.8	95.8	98.8	99.2	9.2	10.0	15.0	17.5	18.3
6	91.3	96.2	98.1	98.1	99.0	49.0	57.7	63.5	66.3	70.2
7	91.7	94.4	97.2	97.2	98.1	12.0	13.0	14.8	16.7	17.3
8	84.8	93.3	94.3	96.2	96.2	44.8	48.6	52.4	56.2	58.1

One interesting point observed from the result is that the Nister & Stewenius approach provides the good results ($> 70\%$ of recalls) on the test sets 1 to 3. In these test sets, the query images consist of only one specific object where the number of instances of a specific object in each query image of these test sets are 1, 2 and 4, respectively (see Figures 11 and 12).

We observed that the Nister & Stewenius approach performed more accurately when the number of instances of the same object in images is increasing. That is, among these three test sets, the Nister & Stewenius approach performed best on the test set 3 (there are 4 instances of the same object in query images). And, it performed worse on the test set 1 in which the query images consist of one instance of an object.

This occurrence can be explained regarding to the key notion that the Bag-of-visual-word image search will search for a single database object whose visual contents without considering spatial information (i.e., collection of visual words) is most similar to the ones extracted from the whole query image. Therefore, as the number of identical object instances is increasing, the visual contents in the query image will be boosted with more relevant visual words extracted from incremental object instances. In contrast to the state-of-the-art proposed in [14], our proposed algorithm can solve the issues of both multiple specific objects and multiple instances of the same object.

6.6. Failure cases

Our proposed algorithm can work very effectively for detection of multiple planar textured objects as we presented in the previous section. However, there are some cases that our algorithm fails to detect the objects in images. First, our algorithm still produces some false positive results. Some examples of this failure case are shown in Figure 16(a). This failure is due to the nature of *a-contrario* frameworks that are inferior if there is small number of inlier matches involved in the *NFA* computation as mentioned in [31]. Second there are some miss-detections as shown in Figure 16(b) where the dashed circles are plotted to indicate the miss-detections. From our observation, these cases are mostly due to adverse illumination changes (e.g. specular noise or glares at CD cases). Finally, our algorithm completely fails to detect the objects in the presence of significant viewpoint changes as illustrated with some examples in Figure 16(c). As mentioned in [54], the problem could be solved by adopting a visual vocabulary that takes into account of viewpoint changes.

7. Conclusions

We have presented a scalable recognition algorithm for simultaneously identifying the identities and detecting the locations of multiple objects in images. Our approach is extended from Bag-Of-Visual-Word (BOVW) image retrieval by incorporating a novel Hough voting based scoring. We also incorporate an *a-contrario* based decision framework into our greedy based hypothesis verification. This allowed our verification algorithm to work insensitively to any threshold. The evaluation with a large scale object database on a set of test images of CD-covers have shown the promising results of our proposed algorithm.

Some of possible future works can be listed as follows. First, we interest in applying the probabilistic Hough voting framework proposed in [55] into our hough voting scoring. This could make our hypothesis generation to be more robust than the current algorithm that is still based on the traditional non-maximum suppression (NMS). Second, with regard to the applications, it is useful if we could make the algorithm to be robust to viewpoint changes. This problem is also central to the works on BOVW large scale image search.

References

[1] R. Szeliski, Computer Vision: Algorithms and Applications, 1st Edition, Springer-Verlag New York, Inc., New York, NY, USA, 2010.

[2] S. S. Tsai, D. M. Chen, V. Chandrasekhar, G. Takacs, N.-M. Cheung, R. Vedantham, R. Grzeszczuk, B. Girod, Mobile product recognition, in: Proceedings of the international conference on Multimedia, MM '10, 2010, pp. 1587–1590.

[3] T. Adamek, D. Marimon, Large-scale visual search based on voting in reduced pose space with application to mobile search and video collections, in: 2011 IEEE International Conference on Multimedia and Expo (ICME), 2011, pp. 1–4.

[4] T. Yeh, J. J. Lee, T. Darrell, Fast concurrent object localization and recognition, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 280 –287.

[5] D. M. Chen, S. S. Tsai, B. Girod, C.-H. Hsu, K.-H. Kim, J. P. Singh, Building book inventories using smartphones, in: Proceedings of the international conference on Multimedia, MM '10, 2010, pp. 651–654.

[6] F.-E. Lin, Y.-H. Kuo, W. H. Hsu, Multiple object localization by context-aware adaptive window search and search-based object recognition, in: Proceedings of the 19th ACM international conference on Multimedia, MM '11, 2011, pp. 1021–1024.

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

[8] A. Collet Romea, M. Martinez Torres, S. Srinivasa, The moped framework: Object recognition and pose estimation for manipulation, International Journal of Robotics Research 30 (1) (2011) 1284 – 1306.

[9] D. Pangercic, V. Haltakov, M. Beetz, Fast and robust object detection in household environments using vocabulary trees with sift descriptors, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Workshop on Active Semantic Perception and Object Search in the Real World, San Francisco, CA, USA, 2011.

[10] J. Pilet, H. Saito, Virtually augmenting hundreds of real pictures: An approach based on learning, retrieval, and tracking, in: Virtual Reality Conference (VR), 2010 IEEE, 2010, pp. 71 –78.

[11] S. Zickler, M. Veloso, Detection and Localization of Multiple Objects, in: Humanoid Robots, 2006 6th IEEE-RAS International Conference on, 2006, pp. 20–25.

[12] R. D. Kai Welke, Pedram Azad, Detection and Localization of Multiple Objects, in: Humanoid Robots, 2006 6th IEEE-RAS International Conference on, 2006.

[13] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, 2003, pp. 1470 –1477 vol.2.

[14] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, Vol. 2, 2006, pp. 2161 – 2168.

[15] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, 2007, pp. 1 –8.

[16] S. Lim, G. Doretto, J. Rittscher, Object Constellations: Scalable, simultaneous detection and recognition of multiple specific objects, in: Proceedings of the ECCV Workshop on Vision for Cognitive Tasks, 2010.

[17] A. Ramisa, S. Vasudevan, D. Scaramuzza, R. L. de Mántaras, R. Siegwart, A tale of two object recognition methods for mobile robots, in: Proceedings of the 6th international conference on Computer vision systems, ICVS'08, 2008, pp. 353–362.

[18] K.-T. Chen, K. H. Lin, Y.-H. Kuo, Y.-L. Wu, W. H. Hsu, Boosting image object retrieval and indexing by automatically discovered pseudo-objects, J. Visual Communication and Image Representation 21 (8) (2010) 815–825.

[19] E. Murphy-Chutorian, J. Triesch, Shared features for scalable appearance-based object recognition, in: Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1. Seventh IEEE Workshops on, Vol. 1, 2005, pp. 16 –21.

[20] S. Kim, I.-S. Kweon, Scalable representation for 3d object recognition using feature sharing and view clustering, Pattern Recognition 41 (2) (2008) 754–773.

[21] H. Jegou, M. Douze, C. Schmid, Improving bag-of-features for large scale image search, International Journal of Computer Vision 87 (3) (2010) 316–336.

[22] A. Desolneux, L. Moisan, J.-M. Morel, Meaningful alignments, International Journal of Computer Vision 40 (1) (2000) 7–23.

[23] A. Desolneux, L. Moisan, J.-M. Morel, From Gestalt Theory to Image Analysis: A Probabilistic Approach, 1st Edition, 2007.

[24] A. Desolneux, A probabilistic grouping principle to go from pixels to visual structures, in: Proceedings of the 16th IAPR international conference on Discrete geometry for computer imagery, DGCI'11, 2011, pp. 1–12.

[25] G. Aragon-Camarasa, J. P. Siebert, Unsupervised clustering in hough space for recognition of multiple instances of the same object in a cluttered scene., Pattern Recognition Letters (11) 1274–1284.

[26] M. Takagi, H. Fujiyoshi, Traffic sign recognition using sift features, IEEJ Transactions on Electronics, Information and Systems 129 (5) (2009) 824–831.

[27] S. Tangruamsub, K. Takada, O. Hasegawa, 3d object recognition using a voting algorithm in a real-world environment, in: Applications of Computer Vision (WACV), 2011 IEEE Workshop on, 2011, pp. 153 –158.

[28] J. Rabin, J. Delon, Y. Gousseau, L. Moisan, Mac-ransac: a robust algorithm for the recognition of multiple objects, in: the Fifth International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT), 2010.

[29] L. Moisan, B. Stival, A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix, International Journal of Computer Vision 57 (3) (2004) 201–218.

[30] A. Desolneux, L. Moisan, J.-M. Morel, From Gestalt Theory to Image Analysis: A Probabilistic Approach, 1st Edition, Springer Publishing Company, Incorporated, 2007.

[31] L. Moisan, P. Moulon, P. Monasse, Image registration with a contrario ransac variant, `http://www.ipol.im/pub/algo/mmm_orsa_homography/` (2011).

[32] B. Leibe, A. Leonardis, B. Schiele, Robust object detection with interleaved categorization and segmentation, International Journal of Computer Vision 77 (1-3) (2008) 259–289.

[33] K. Mikolajczyk, B. Leibe, B. Schiele, Multiple object class detection with a generative model, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, Vol. 1, 2006, pp. 26 – 36.

[34] N. Razavi, J. Gall, L. Van Gool, Scalable multi-class object detection, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011, pp. 1505 –1512.

[35] D. Das, Y. Kobayashi, Y. Kuno, Multiple object category detection and localization using generative and discriminative models, IEICE Transactions 92-D (10) (2009) 2112–2121.

[36] W.-L. Zhao, X. Wu, C.-W. Ngo, On the annotation of web videos by efficient near-duplicate search, Multimedia, IEEE Transactions on 12 (5) (2010) 448 –461.

[37] H. Xie, K. Gao, Y. Zhang, J. Li, Y. Liu, Pairwise weak geometric consistency for large scale image search, in: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR '11, 2011, pp. 42:1–42:8.

[38] A. A. Paulino, J. Feng, A. K. Jain, Latent fingerprint matching using descriptor-based hough transform, in: Biometrics (IJCB), 2011 International Joint Conference on, 2011, pp. 1 –7.

[39] Z. Lei, T. Fang, H. Huo, D. Li, Rotation-invariant object detection of remotely sensed images based on texton forest and hough voting, Geoscience and Remote Sensing, IEEE Transactions on PP (99) (2011) 1 –12.

[40] H. Jegou, M. Douze, C. Schmid, On the burstiness of visual elements, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 1169 –1176.

[41] R. I. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, 2nd Edition, Cambridge University Press, ISBN: 0521540518, 2004.

[42] A. Desolneux, L. Moisan, J.-M. Morel, A grouping principle and four applications, IEEE Trans. Pattern Anal. Mach. Intell. 25 (4) (2003) 508–513.

[43] R. G. von Gioi, J. Jakubowicz, J.-M. Morel, G. Randall, Lsd: A fast line segment detector with a false detection control, IEEE Trans. Pattern Anal. Mach. Intell. 32 (4) (2010) 722–732.

[44] T. Veit, F. Cao, P. Bouthemy, An *a contrario* decision framework for region-based motion detection, International Journal of Computer Vision 68 (2) (2006) 163–178.

[45] A. Robin, L. Moisan, S. Le Hegarat-Mascle, An a-contrario approach for subpixel change detection in satellite imagery, Pattern Analysis and Machine Intelligence, IEEE Transactions on 32 (11) (2010) 1977 –1993.

[46] P. Musé, F. Sur, F. Cao, Y. Gousseau, J.-M. Morel, An *a contrario* decision method for shape element recognition, International Journal of Computer Vision 69 (3) (2006) 295–315.

[47] N. Sabater, A. Almansa, J. Morel, Meaningful matches in stereovision, Pattern Analysis and Machine Intelligence, IEEE Transactions on PP (99) (2011) 1.

[48] D. Monnin, E. Bieber, G. Schmitt, A. L. Schneider, An effective rigidity constraint for improving ransac in homography estimation, 2010.

[49] A. Vedaldi, B. Fulkerson, VLFeat: An open and portable library of computer vision algorithms, `http://www.vlfeat.org/` (2008).

[50] F. Fraundorfer, C. Wu, J.-M. Frahm, M. Pollefeys, Visual word based location recognition in 3d models using distance augmented weighting, in: Fourth International Symposium on 3D Data Processing, Visualization and Transmission, 2008.

[51] M. Aly, P. Welinder, M. Munich, P. Perona, Towards Automated Large Scale Discovery of Image Families, in: Second IEEE Workshop on Internet Vision,IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[52] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International Journal of Computer Vision 88 (2) (2010) 303–338.

[53] N. Ichimura, Recognizing multiple billboard advertisements in videos., in: 2006 IEEE Pacific-Rim Symposium on Image and Video Technology (PSIVT), 2006, pp. 463–473.

[54] D. Chena, S. S. Tsaia, V. Ch, G. Takacsa, J. S. A, B. Girod, Robust image retrieval using multiview scalable vocabulary trees, in: Proceedings of Visual Communication and Image Processing, 2009, 2009.

[55] O. Barinova, V. Lempitsky, P. Kohli, On detection of multiple object instances using hough transforms, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, 2010, pp. 2233 –2240.

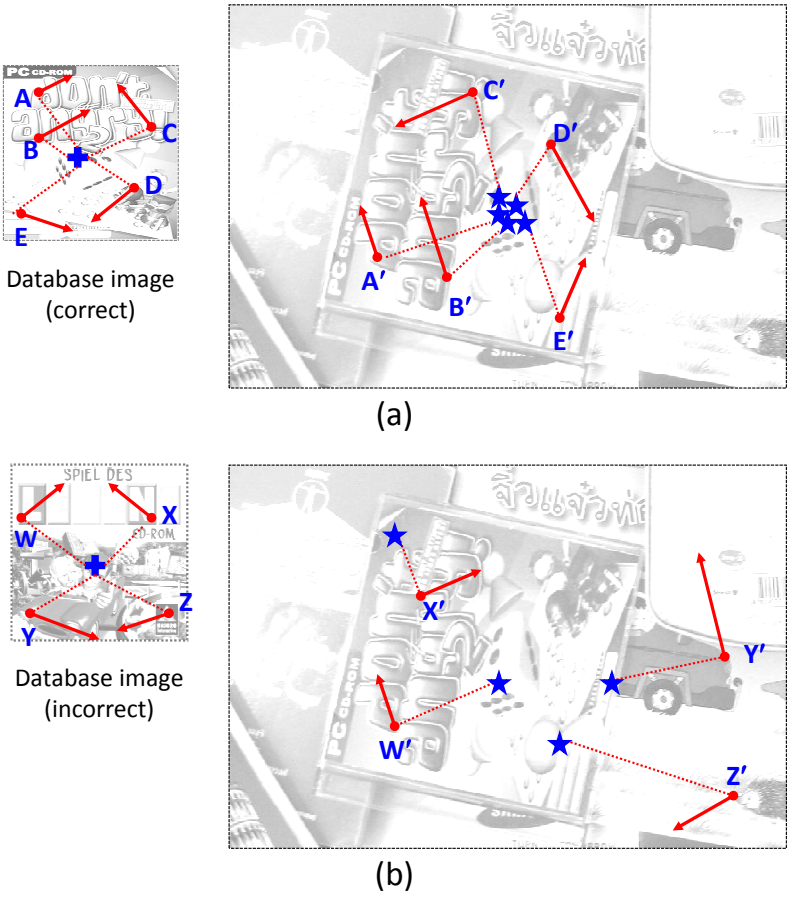


Figure 5: Illustration examples of voting (a) Voted locations (star markers) from correct matches $A-A'$ to $E-E'$. They tend to cluster at some location. (b) Voted locations from incorrect matches $W-W'$ to $Z-Z'$. Unlike (a), the voted locations are not clustered.

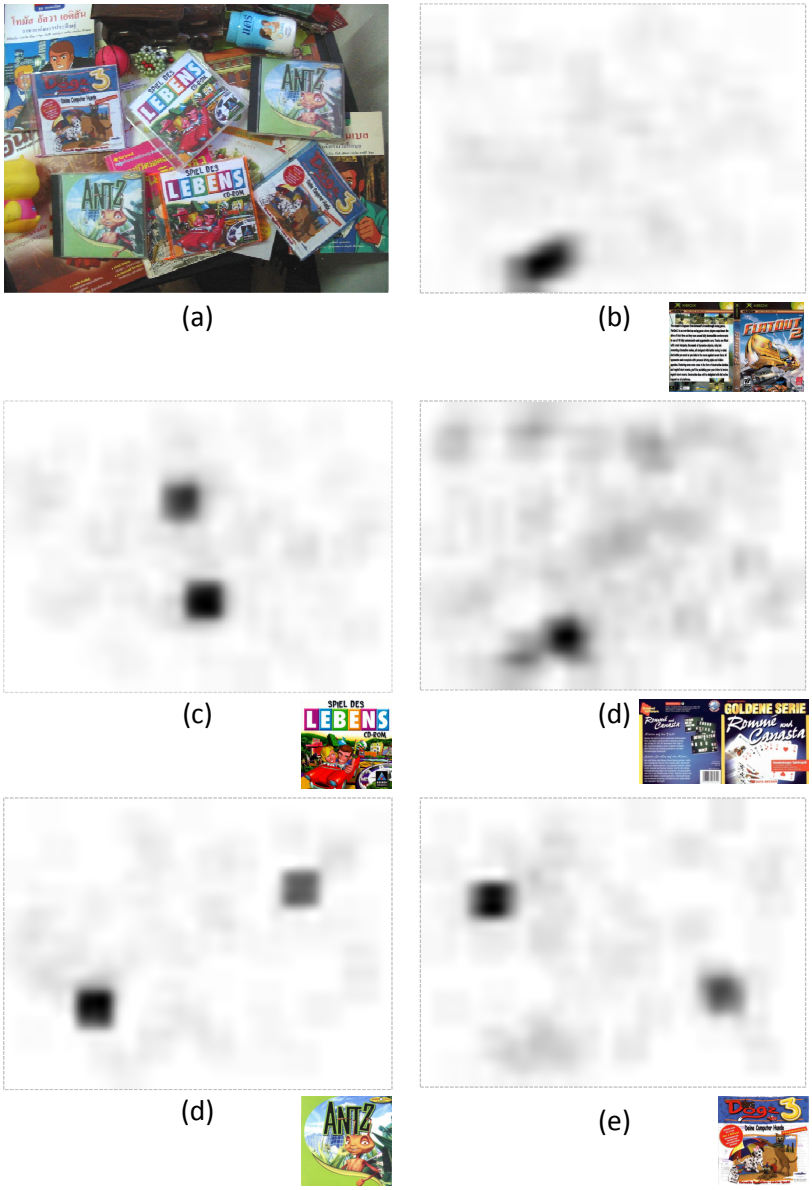


Figure 6: Example of voted spaces computed from the test image in Figure 1. (a) Test image. (b) to (e) are the voted spaces for the top 5 object retrieval results from our Hough voting based scoring. The database images of retrieved objects are attached on the below of illustrated voted space images. Note that these top 5 objects are similar to the ones in Figure 1-c

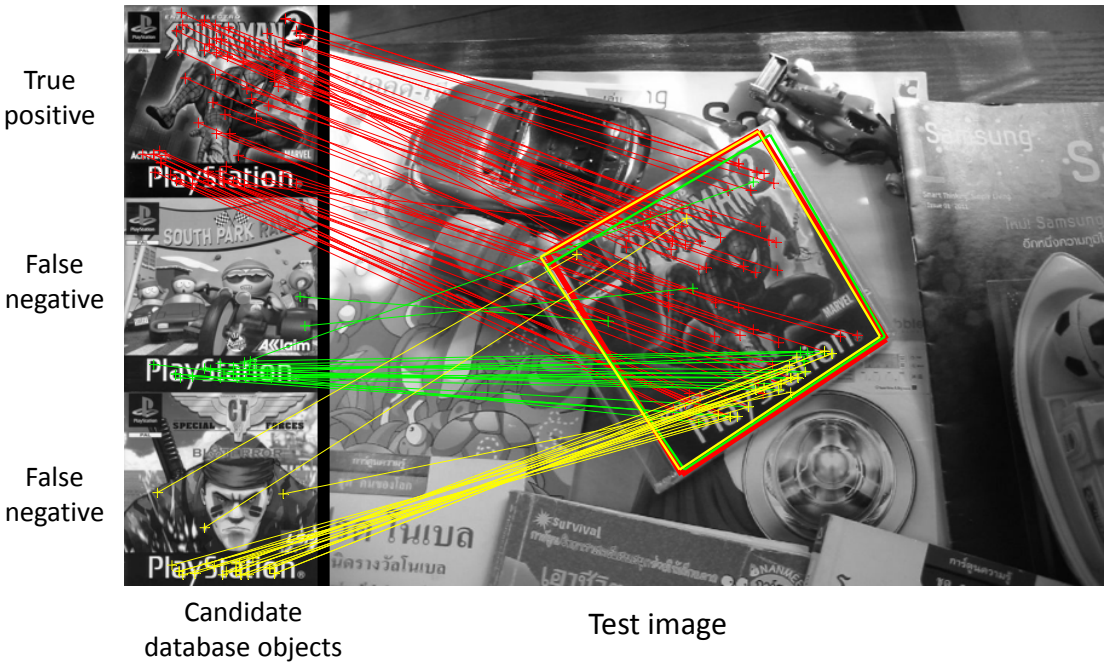


Figure 7: Example of conflicting hypotheses.

Algorithm 1 Hypothesis verification

Input: $H = \{h_k; k = 1, 2, ..\}$: A set of hypotheses.
where $h_k = \{(o_k, P_k)\}$: o_k=object identity, P_k=putative matches.
where $P_k = \{(p_{k,i}, p'_{k,i})\}$: $p_{k,i}, p'_{k,i}$ are database keypoint and test image keypoint, respectively.
Output: $L = \{(o_j, BB_j); j = 1, 2, ..\}$: Detection result.
where o_j=object identity, BB_j=object bounding box.
Parameter: ϵ: Hypothesis quality threshold (fixed to 1.0)

$L \leftarrow \phi$
stop $\leftarrow$ **false**
while *stop* $\neq$ **true do**
  // Determine the inlier matches and hypothesis scores
  // of remaining hypotheses
  for $i = 1$ to $|H|$ **do**
    $(S_i, \mathbf{A}_i, S'_i) \leftarrow$ **RansacPlanarTransfEstimation**(P_i)
    $nfa \leftarrow$ **ComputeNFA**$(S_i, \mathbf{A}_i, S'_i, P_i)$
    $Q(h_i) \leftarrow -\log(nfa)$
  end for

  // Find the best available hypothesis
  Find j that $\max_{1 \leq j \leq |H|} Q(h_j)$

  // If the quality score $\geq -\log(\epsilon)$ *then accept the hypothesis*
  // remove all matches of other remaining hypotheses that are
  // associated with the test image keypoints in
  // the bounding box of the accepted hypothesis
  if $Q(h_j) \geq -\log(\epsilon)$ **then**
    $DbObjBB \leftarrow$ {Bounding box vertices of the database object o_j}
    $BB_j \leftarrow$ **PlanarTransfProjection**$(\mathbf{A}_j, DbObjBB)$
    $L \leftarrow L \cup \{(o_j, BB_j)\}$
    $H \leftarrow H \backslash \{h_j\}$
    for $i = 1$ to $|H|$ **do**
      $P_i \leftarrow$ **RemoveMatch**(P_i, BB_j)
    end for
  else
    stop $\leftarrow$ **true**
  end if
end while

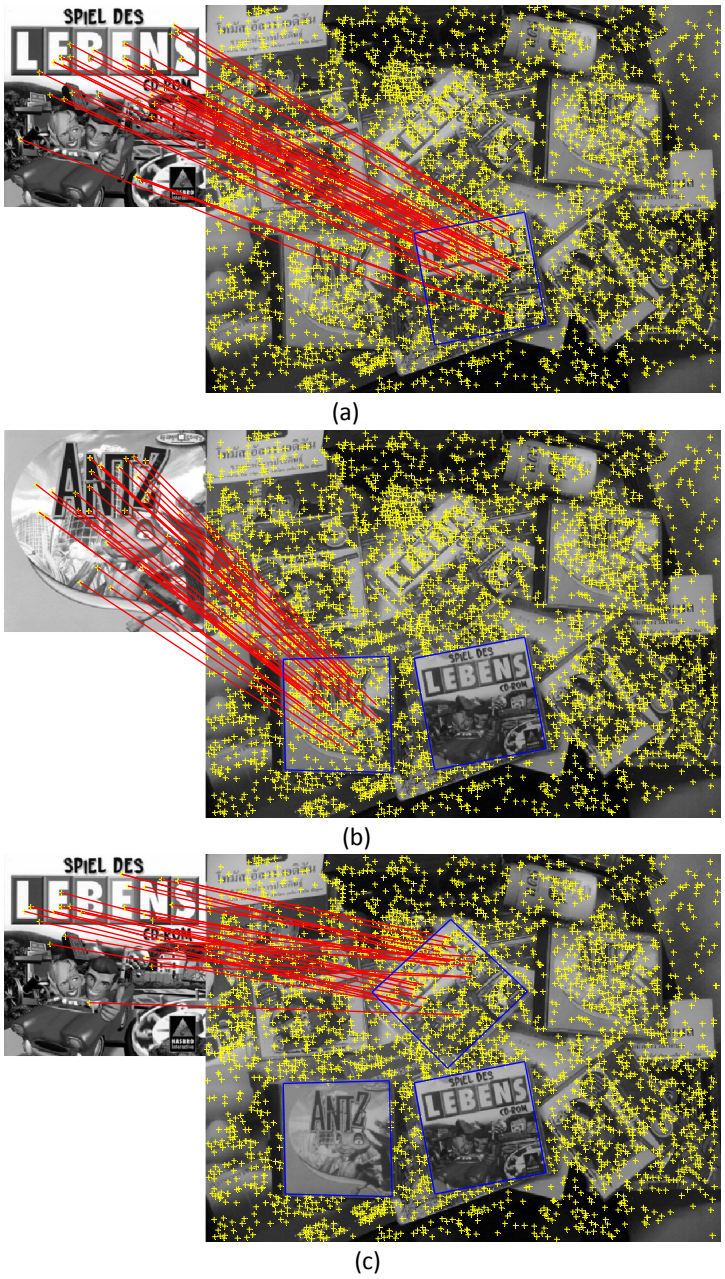
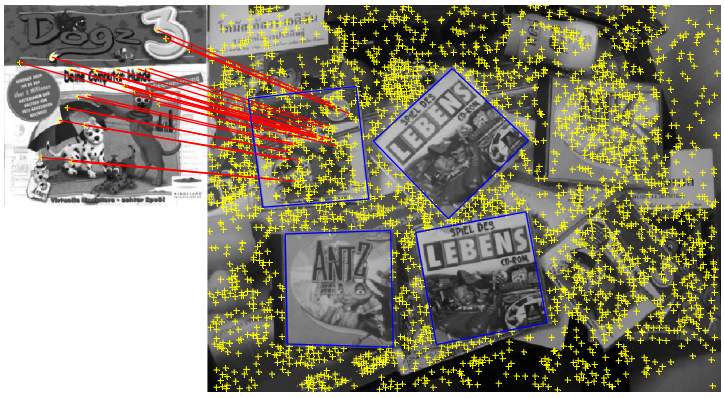


Figure 8: Iterations of hypothesis verification: (a) to (c) for iteration 1 to 3 respectively. The left (smaller) images are the database object images. The right image is the test image. The yellow markers show the extracted keypoints. The red lines show the matches between the test image keypoints and the database keypoints. The blue bounding boxes show the detected instances.



(d)

(e)

(f)

Figure 8: (continue) Iterations of hypothesis verification: (d) to (f) for iteration 4 to 6 respectively.

Figure 9: Some examples of training images of CD-covers.

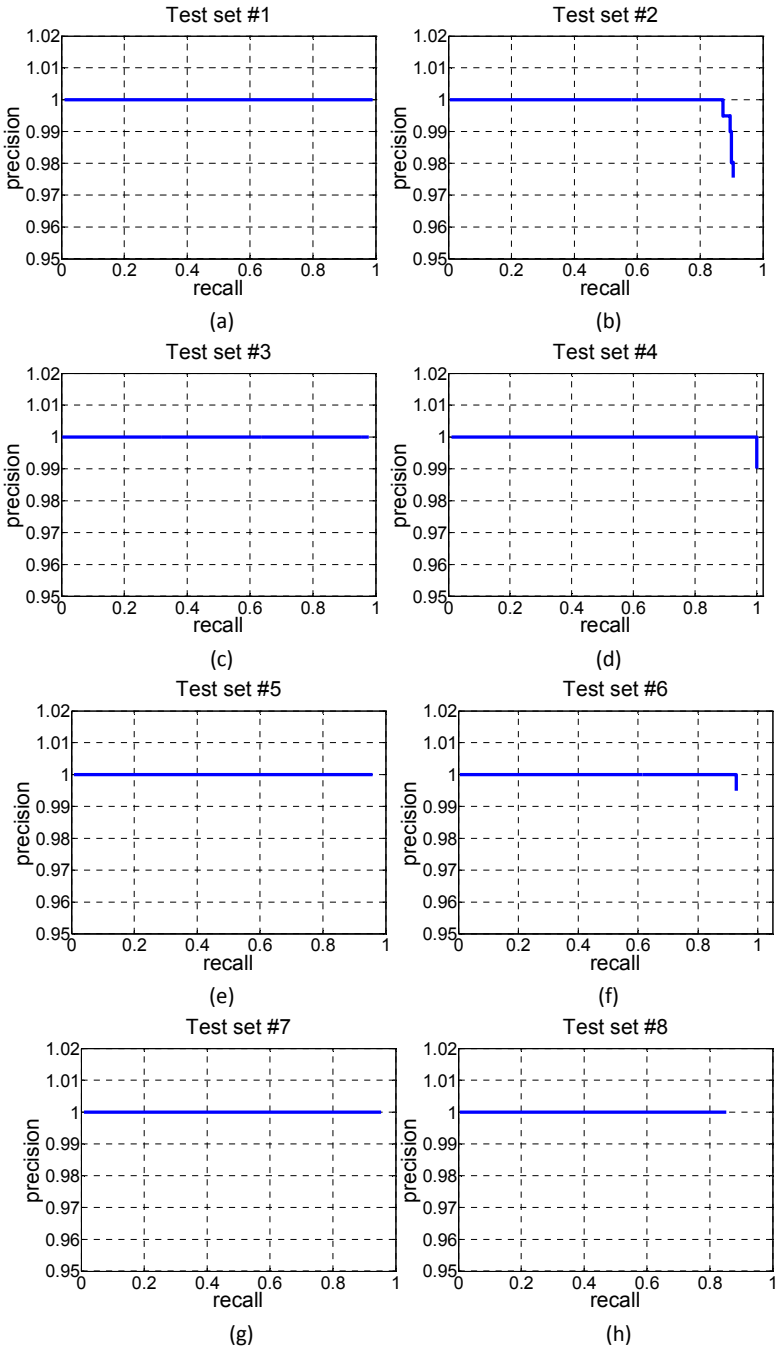


Figure 10: Precision/Recall curves of the data-sets. Figures (a) to (h) for the test sets 1 to 8, respectively.

Figure 11: Some qualitative results on the test set 1 (first two rows) and 2 (last two rows). The bounding boxes of detected objects are shown in red.

Figure 12: Some qualitative results on the test set 3 (first two rows) and 4 (last two rows).

Figure 13: Some qualitative results on the test set 5 (first two rows) and 6 (last two rows).

Figure 14: Some qualitative results on the test set 7 (first two rows) and 8 (last two rows).

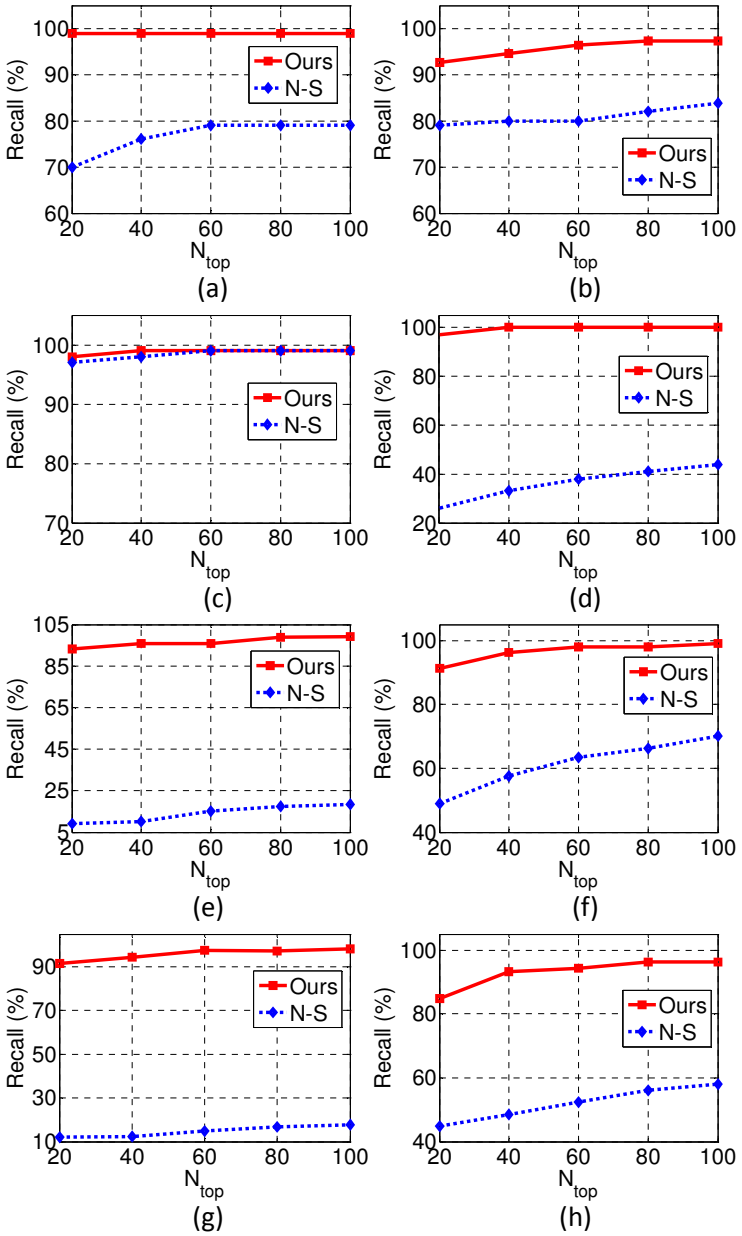


Figure 15: Comparative results between our proposed ranking scheme (in solid red) and Nister & Stewenius (denoted N-S) [14] (in dash blue). Figures (a) to (h) are corresponding for the test sets 1 to 8, respectively. The y-axis shows the recall in percentage. The x-axis shows the number of top candidate objects in the recall evaluation.

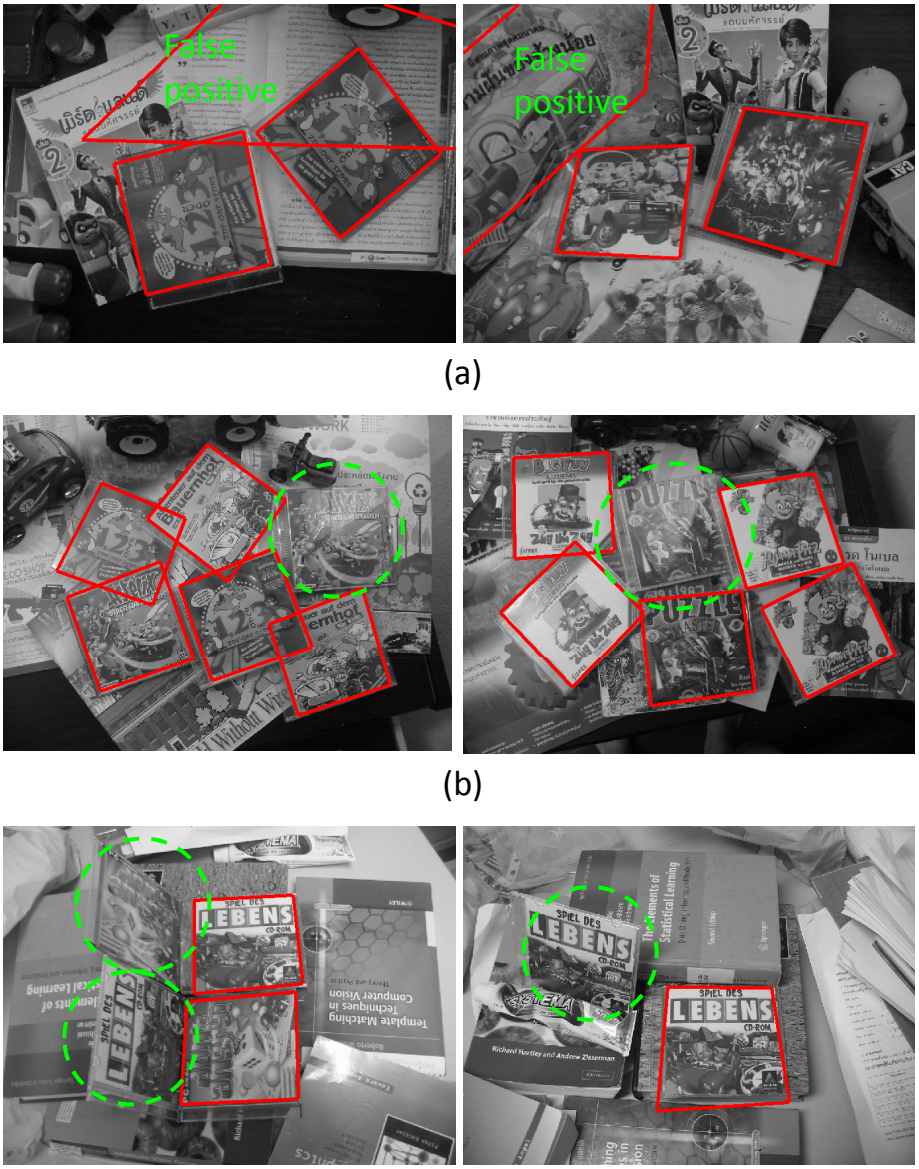


Figure 16: Example of failure cases: (a) False positives. (b) miss detections due to adverse illumination changes (glares at CD cases). (c) miss detections due to viewpoint changes. Note that the green dashed circles indicate undetected instances.

หนังสือส่งถึงบรรณาธิการ เพื่อส่งวารสารไปตีพิมพ์

Srinakharinwirot University
114 Sukhumvit 23,
Bangkok 10110, Thailand

February 6, 2012

Dear Professor Avinash C. Kak
Editor-in-Chief, CVIU

Please consider our submission for publication in Computer Vision and Image Understanding.

Some of our graphical plots include color for better discrimination between the curves. Should our paper be accepted and if the editors so desire, we will be glad to redo those plots in black and white.

Please note that this work has not been submitted for publication elsewhere.

Sincerely,

Pradit Mittrapiyanuruk
praditm@swu.ac.th
Srinakharinwirot University

From: pakorn.kae@kmutt.ac.th
Subject: [Fwd: Computer Vision and Image Understanding: Submission Confirmation]
Date: Mon, February 6, 2012 6:40 pm
To: "Pradit" <praditm@swu.ac.th>

---------- Original Message ----------
Subject: Computer Vision and Image Understanding: Submission Confirmation
From: "CVIU (ELS)" <cviu@elsevier.com>
Date: Mon, February 6, 2012 6:16 pm
To: pakorn.kae@kmutt.ac.th

Title: Scalable Detection of Multiple Specific Objects using
Bag-Of-Visual-Word Image Retrieval with Hough Voting based Ranking
Corresponding Author: Dr. Pakorn Kaewtrakulpong
Authors: Pradit Mittrapiyanuruk, Ph.D.;

Dear Dr. Kaewtrakulpong,

This is to confirm that the above-mentioned manuscript has been received
for consideration in Computer Vision and Image Understanding.

You will be able to check on the progress of your manuscript by logging on
to the Elsevier Editorial System for Computer Vision and Image
Understanding as an author:
http://ees.elsevier.com/cviu/
Your username is: pakorn
If you need to retrieve password details, please go to:
http://ees.elsevier.com/cviu/automail_query.asp

Your paper will be assigned a manuscript number shortly and you will soon
receive an e-mail with this number for your reference.

Thank you for submitting your manuscript to Computer Vision and Image
Understanding. Should you have any questions, please feel free to contact
our office.

Kind regards,

Linda Shapiro
Journal Manager
Computer Vision and Image Understanding
cviu@elsevier.com

For further assistance, please visit our customer support site at
http://support.elsevier.com Here you can search for solutions on a range
of topics, find answers to frequently asked questions and learn more about
EES via interactive tutorials. You will also find our 24/7 support contact
details should you need any further assistance from one of our customer
support representatives.

--
Computer Center staff will *NEVER* send you email requesting your password
information. Please ignore any email messages that claim to require you to provide
such information

From: pakorn.kae@kmutt.ac.th
Subject: [Fwd: Computer Vision and Image Understanding Submission: Manuscript Number Assigned]
Date: Thu, February 9, 2012 7:17 pm
To: "Pradit" <praditm@swu.ac.th>

---------- Original Message ----------
Subject: Computer Vision and Image Understanding Submission: Manuscript Number Assigned
From: "CVIU (ELS)" <cviu@elsevier.com>
Date: Thu, February 9, 2012 5:07 pm
To: pakorn.kae@kmutt.ac.th

Ms. No.: CVIU-12-54
Title: Scalable Detection of Multiple Specific Objects using Bag-Of-Visual-Word Image Retrieval with Hough Voting based Ranking
Corresponding Author: Dr. Pakorn Kaewtrakulpong
Authors: Pradit Mittrapiyanuruk, Ph.D.;

Dear Dr. Kaewtrakulpong,

Your submission, referenced above, has been assigned the following manuscript number: CVIU-12-54

You will be able to check on the progress of your paper by logging on to the Elsevier Editorial System as an author:
http://ees.elsevier.com/cviu/
Your username is: pakorn
If you need to retrieve password details, please go to:
http://ees.elsevier.com/cviu/automail_query.asp

Thank you for submitting your work to Computer Vision and Image Understanding.

Kind regards,

Linda Shapiro
Journal Manager
Computer Vision and Image Understanding
cviu@elsevier.com

Please take into account that Electronic Annexes can be appended to your document, and archived on the web site of the journal. They can include animations, video or audio clips, demos, additional data, etc. These electronic annexes are free of charge. You can find a guide for multimedia files at
http://www.elsevier.com/wps/find/authorsview.authors/movies_animations.

For technical questions regarding text, figures or video file formats, you may also contact the Journal Manager of CVIU, Linda Shapiro,
CVIU@elsevier.com

****

For guidelines on how to track your manuscript in EES please go the following address: http://support.elsevier.com/app/answers/detail/a_id/89

--
Computer Center staff will *NEVER* send you email requesting your password information. Please ignore any email messages that claim to require you to provide

such information

Submissions Being Processed for Author Pakorn Kaewtrakulpong, Ph.D.

Page: 1 of 1 (1 total submissions)

Action	Manuscript Number	Title
View Submission Send E-mail	CVIU-12-54	Scalable Detection of Multiple Specific Objects using Bag-Of-Visual-Wor Hough Voting based Ranking

Page: 1 of 1 (1 total submissions)

<< Author Main Menu