UNIVERSITY OF CALIFORNIA
RIVERSIDE

Exploring Semantic Concept Co-Occurrences for Image Based Applications

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

by

Linan Feng

December 2014

Dissertation Committee:

Dr. Bir Bhanu, Chairperson
Dr. Chinya Ravishankar
Dr. Tao Jiang
Dr. Vagelis Hristidis

The Dissertation of Linan Feng is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

Upon the completion of this work, I own my gratitude to a great number of people. I would first like to express my deepest gratitude to Dr. Bir Bhanu, for his support and guidance as my advisor during my PhD study. I would like to thank my committee members, Dr. Chinya Ravishankar, Dr. Tao Jiang and Dr. Vagelis Hristidis for their constructive comments to improve this work. I would also like to thank Dr. Marek Chrobak and Dr. Vassilis Tsotras who were in my oral qualifying exam committee and gave me insightful feedback and suggestions.

I am grateful to Zhixing Jin, Xiaojing Chen and Dr. Ninad Thakoor, with whom I collaborated extensively and shared a lot of inspirational ideas. I am also grateful to Suresh Kumar, as my colleague and friend, who had brought enormous optimism to all of us. I am indebted and grateful to my wife, Yimei Zhang, for her companion and support to my research and all other aspects in my life.

I would like to thank Dr. Yu Sun, Dr. Le An, Dr. Alberto Cruz and all of my other current and former colleagues, friends at University of California, Riverside and at other places who have offered me help, support, and joy. I would also like to thank all of my family members who have been supportive in many aspects during the completion of this work.

The materials of some chapters in this dissertation have appeared in "Understanding Dynamic Social Grouping Behaviors of Pedestrians" © 2014 IEEE by L. Feng *et al.*. This work has been supported by US National Science Foundation (NSF) grants 0641076 and 0727129, NSF grants 0905671 and 1330110 and ONR grant N00014-12-1-1026.

I dedicate this Dissertation,

To my father, Ruiqiang Feng, and my mother, Yanfeng Wei,

for their unconditional love, support, sacrifice, and encouragement.

To my wife, Yimei Zhang, for her support, understanding, and endless love.

Without you, I would not have gone so far.

ABSTRACT OF THE DISSERTATION

Exploring Semantic Concept Co-Occurrences for Image Based Applications

by

Linan Feng

Doctor of Philosophy, Graduate Program in Computer Science
University of California, Riverside, December 2014
Dr. Bir Bhanu, Chairperson

Describing visual image contents by semantic concepts is an effective and straightforward way to facilitate various high level applications. Inferring semantic concepts from low-level pictorial feature analysis is challenging due to the semantic gap problem, while manually labeling concepts is unwise because of a large number of images in both online and offline collections. In this paper, we present a novel approach to automatically generate intermediate image descriptors by exploiting concept co-occurrence patterns in the pre-labeled training set that renders it possible to depict complex scene images semantically. Our work is motivated by the fact that multiple concepts that frequently co-occur across images form patterns which could provide contextual cues for individual concept inference. We discover the co-occurrence patterns as hierarchical communities by graph modularity maximization in a network with nodes and edges representing concepts and co-occurrence relationships separately. A random walk process working on the inferred concept probabilities with the discovered co-occurrence patterns is applied to acquire the refined high level image semantic representation. Through experiments in applications including automatic image annotation, semantic image retrieval, moth species identification and multi-pedestrian tracking on several challenging datasets, we demonstrate the effectiveness of the proposed concept co-occurrence patterns as well as the proposed image semantic representation in comparison with state-of-the-art approaches.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Representing images by semantic concepts instead of visual features remains a challenging problem. Generating semantic descriptors manually is not feasible due to the ever-growing number of image collections. Current machine intelligence and statistical learning techniques for inferring semantic concepts from low-level features struggle in bridging the semantic gap [104]. However, many image-based applications such as retrieval, annotation, recommendation, indexing and ranking, require an effective semantic representation of images. There is a growing need in automatically inferring concepts from visual properties by learning the correspondence from loosely labeled data.

Semantic concepts cover not only objects that are used in many recognition tasks but also topics at the semantic levels beyond single objects. These higher semantic level could be a scene (e.g., beach), an event (e.g., commencement), and a piece of knowledge (e.g., how to drive a car). A simple form of contextual information is the co-occurrence frequencies of groups of concepts that appear across images with similar scenes. Visual co-occurrence can be quite important in providing semantic cues in inferring concepts compared to other conceptual and perceptual models [160] such as the WordNet distance [136] which is built upon semantic similarity. It has been shown [62] that co-occurrence of concepts could consolidate the appearance of each concept in an image. For example, if “horse” and “windmill” forms a co-occurrence pattern, then the probability of occurrence of “horse” could be reinforced by a strong confidence of “windmill” inference, while the occurrence of “zebra” could be rejected because it has a weak co-occurrence with “windmill”. Discovering co-occurrence patterns of semantic concepts is an essential step to encode contextual information into the individual concept inference.

In Chapter 3, we propose a novel approach to discover the co-occurrence patterns in a network structure where the nodes represent semantic concepts and edges represent co-occurrences. The significance of the co-occurrence relationship between two concepts is denoted by the edge weight. A common property that has been discovered in many networks is the *community structure* property, which is the partition of network nodes into groups (communities) with highly inter-connected nodes (more edges with higher weights), and nodes belonging to different groups being sparsely connected (fewer edges with lower weights). Inspired by the theories in network analysis, we discover the concept co-occurrence patterns by identifying communities in a network. We adopt modularity optimization [118] based approach to uncover *hierarchical community structure* which naturally reflects the co-occurrence patterns at different closeness levels. The idea of hierarchical community structure and co-occurrence patterns is illustrated in Figure 1.1. To our knowledge, our work is the first attempt to explore concept co-occurrences from the network analysis point of view.

In Chapter 3, we also introduce a novel random walk based approach to utilize the discovered co-occurrence patterns to generate “concept signature”, a new image

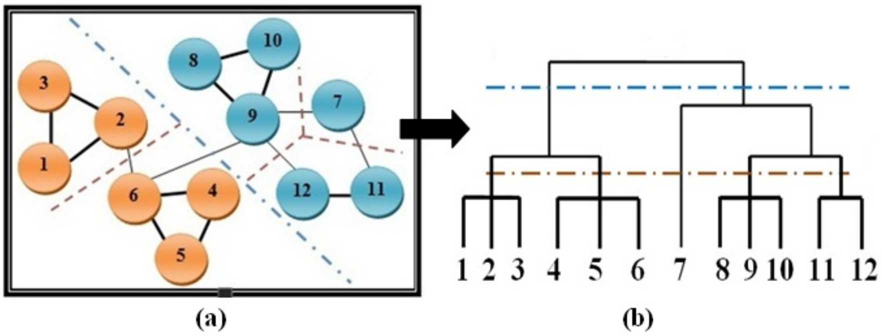


Figure 1.1: An illustration of (a) a network of nodes representing the semantic concepts and the edges representing the co-occurrence relations, and (b) the discovered corresponding hierarchical community structure from the network that shows concept co-occurrence patterns at different levels.

representation using high-level semantic concepts to assist in image annotation and retrieval. The hypothesis here is that the probability scores of uncertain semantic concepts in the concept signature that are generated from the inference model can be promoted or weakened based on the reliably inferred members in a co-occurrence pattern. We demonstrate that our concept signature representation can be very useful in annotation and retrieval of complex scene images. Experimental results in image annotation and retrieval application scenarios on popular benchmark datasets show clear gains from co-occurrence patterns as compared to other baseline approaches with/without exploiting concept correlations.

Manually collecting, identifying, archiving and retrieving specimen images is an expensive and time-consuming work for entomologists. There is a clear need to introduce fast computer systems integrated with modern image processing and analysis algorithms to accelerate the process. In Chapter 4, we describe the development of an automated moth species identification and retrieval system (SPIR) using computer vision and pattern recognition techniques. The core of the system is a probabilistic model that infers Semantically Related Visual (SRV) attributes from low-level visual features of moth images in the training set, where moth wings are segmented into information-rich patches from which the local features are extracted, and the SRV attributes are provided by human experts as ground-truth. For the large amount of unlabeled testing images in the database or added into the database later on, an automated identification process is evoked to translate the detected salient regions of low-level visual features on the moth wings into meaningful semantic SRV attributes. We further propose a novel network analysis based approach to explore and utilize the co-occurrence patterns of SRV attributes as contextual cues to improve individual attribute detection accuracy. The effectiveness of the proposed approach is evaluated in automated moth identification and attribute-based image retrieval. In addition, a novel image descriptor called SRV attribute signature is introduced to record the visual and semantic properties of an image and is used to compare image similarity. Experiments are performed on an existing entomology database to illustrate the capabilities of our proposed system. We observed that the system performance is improved by the SRV attribute representation and their co-occurrence patterns.

Additionally, we apply the idea of co-occurrence pattern detection in a very interesting application which is known as group-based multi-pedestrian tracking using single camera. The details are discussed in Chapter 5. Consider a video clip recording a number of pedestrians walking in an outdoor (indoor) environment such as a square (hall). Imagine an algorithm that is able to analyze the video and answer the questions like: Are these people evacuating from an emergent situation? Are they gathering for a special event? By just looking at each individual it could be very hard to train the computers to understand these high-level concepts from the low-level visual representations. In Chapter 5, we introduce a new model for analyzing social behaviors among pedestrians: rather than treating each person in isolation, we analyze their social grouping behaviors so as to reinforce the recognition of movements of each individual in a group. Our approach is inspired by recent achievements in computer vision and pattern recognition where the correlations of semantic or geometrical concepts are utilized as extra contextual information for recognizing objects in complex scenes [57]. In our work, pedestrian detection and interactions are enforced by taking the advantage of contextual information that comes from within-group positional, velocity and directional distance consistences. This provides our approach the robustness to pedestrian walking behavior analysis from dynamic cluttered background, occlusions among pedestrians, illumination and viewpoint changes, or the variations of backgrounds caused by mobile cameras such as smart-phones.

Each chapter in this dissertation stands alone as a complete description of each aforementioned method and application. Before we dive into details of individual methods, related work is presented in Chapter 2.

Chapter 2

Related Work

In the following, we review those approaches that are most relevant to our research along five directions: **(i)** Models that investigate concept correlations as contextual information for image based applications, **(ii)** Image semantic descriptors, **(iii)** Network analysis approaches for detecting communities (co-occurrence patterns), **(iv)** Automated insect identification and image retrieval systems and **(v)** crowd scene analysis and multi-people tracking.

2.1 Semantic Concept Co-occurrence Models

The approaches based on co-occurrence models for concept inference in complex scene images have gained an increasing popularity [38, 79, 158]. In [62, 136], pairwise concept co-occurrence has been integrated into the concept categorization framework by using a co-occurrence matrix. These approaches have several advantages over standard concept inference techniques, for example, incorporating semantic context compensates the ambiguity of concept visual appearance. However, the matrix of the co-occurrence has an inevitable pairwise constraint on the relationship.

Several recent works explore multi-concept learning/detection techniques for automated image annotation that aim to model the co-occurrence information among concepts/annotations. A simple way is to rank the related concepts based on their co-occurrence relations in the training set and use the ranked relations to refine the annotation results. The idea is similar to collaborative filtering (CF) [70] used by the recommender systems [47]. CF has been introduced in image retrieval [163] to collect the relevance feedback co-occurrences. One of the challenges for CF is the data sparsity problem where the image-concept matrix used for collaborative filtering could be extremely large and sparse in a large image dataset. Matrix-factorization (MF) [103] has been found to be accurate and scalable to address the sparsity problem in CF. By introducing the non-negative constraint into the MF process (NNMF), Zhou et al. [198] proposed a CF method for concept correlation estimation, and Liu et al. [101] presented a framework for semi-supervised multi-label learning using NNMF. Li et al. [98] proposed a multi-correlation probabilistic matrix factorization model to seamlessly estimate the image-concept, image-image and concept-concept correlations simultaneously. Desai et al. [34] examine spatial co-occurrence statistics and incorporate it as contextual relations. Our approach in this paper is significantly different from the above works in discovering the co-occurrence patterns of concepts of any size by detecting the patterns as social communities in a network structure.

To learn more reliable contextual relationships among the semantic concepts, multi-task learning [48] has been introduced for hierarchical image annotation which requires the incorporation of concept ontology. Fan et al. [49] constructed the concept ontology using semantic and visual similarity of concepts, in an attempt to explore the inter-concept correlations and to organize the image concepts in a hierarchy. Multi-task learning is adopted to overcome the problem of intra-concept visual variations. Bourdev et al. [50] presented a hierarchical concept learning framework by incorporating concept ontology and multi-task learning to enhance the image classification performance with a large concept vocabulary. Our approach not only avoids the pairwise constraint, but also, more interestingly, it relies more on the contextual relationships (co-occurrences) rather than the perceptual relationships (concept ontology) that are used in the multi-task learning frameworks [48, 49, 50].

Another problem in existing approaches is that one concept cannot be shared among co-occurrence groups. For example, the method proposed in [27] attempts to discover the co-occurrence between objects by learning a tree structure using Chow-Liu algorithm based on pairwise mutual information. But in their tree structure a concept at the root can only have relationships with the children in its subtree, and cannot have any relationship with the nodes in its siblings' subtrees. Also the same concept cannot be duplicated and shared between subtrees. For instance in their tree structure, sky only has a connection with mountain but not with tree and road which may not be true in many cases. In contrast, our proposed approach addresses the overlapping of concepts explicitly.

One of the drawbacks in existing work [136, 160] is the dataset limitation. To find the co-occurrence relationships between objects, these papers do not use strongly labeled data. Instead, they rely on outside sources such as Google Sets, WordNet and Word Association. However, these sources usually do not consider the visual co-occurrence, namely, they are purely based on text or semantic meaning similarity. For example, Google Sets leverage the word co-occurrence on web pages *without* considering the actual observations in images. WordNet is purely based on the semantic meaning similarity to determine the distance between concepts. It does not reflect the actual co-occurrence property in images. However, in our work, we use the datasets for which the labels are given only when the corresponding concept are observed in an image.

Many algorithms for detecting concept correlations used graph models which is close to our idea. Probabilistic graph models that focus on batch-mode concept detection are proposed in [178]. Correlation of concept co-occurrence and relative spatial locations in images are captured by a tree model in [27]. Besides the positive correlations, they also modeled the negative relationships in the tree structure.

2.2 Image Semantic Descriptors

Many papers in computer vision adopt semantic representations for multimedia understanding and scene analysis, and for applications such as semantic based image annotation and retrieval [41, 148, 186]. Berg et al. [165] automatically generate natural language sentences from *gist* features at different image sizes. Since their final goal is to generate sentence description for an image, the image descriptor is only used in an intermediate procedure and it is still based on visual features which will have a gap between the semantic meaning of images. Unlike these descriptors, our image signature representation focuses on mid-level semantic concepts that are not too general (e.g., *forest, desert*) and not too specific (e.g., *palm tree, NIKE shoes*) and addresses the semantic gap problem explicitly.

Farhadi et al. [3] generate semantic descriptions for images in the form of sentence annotations. Instead of predicting sentence from an image directly, they provide an intermediate step to compute the meaningful triplet (object, action and scene). They name the set of triplets as *meaning space.* The idea of finding the most matched triplet from the meaning space for an image is similar to our concept of finding co-occurrence patterns from the network structure. However, the meaning space is used only as an intermediate step for predicting the sentences, it is not used as a semantic descriptor for comparing image similarity as in our work.

Attribute representation has become a trend in image classification [95, 14] and visual recognition [53, 91] due to its intuitive way in interpreting images and cross-category generalization [51]. Unlike visual words, semantic attributes are sharable discriminative visual properties that are machine-detectable and human nameable (e.g., "square" as a shape property, "silk" as a texture property, "has wing" as a sub-component property, and "can fly" as a functional property). One advantage of semantic attributes is that they naturally bridge the gap between low-level visual features and high-level concepts. In other words, semantic attributes can be used to answer not only "how" two images are similar in a human interpretable way [173], but also "why" an image is identified to belong to a specific category [126]. Attributes are also used frequently in multimedia retrieval as an intermediate semantic description [42]. As compared to the attribute-based representations, our concept signature is generated from the inference models combined with a refinement process that utilizes the co-occurrence information of the concepts.

The most similar image descriptor to our concept signature is the *Object Bank* representation [98]. However, there are several differences. First, the Object Bank representation is computed on grids over an entire image but the grids usually do not fully match the object geometry. Instead, we compute concept signature for each segmented salient region and the signatures are concatenated to form the final image descriptor. Second, each object in the bank is selected based on the occurrence frequency across different datasets. However, we do not consider cross-dataset concept occurrence because an indoor concept may not have frequent occurrences in an outdoor dataset. Third, object bank is used to address the scene classification and object recognition tasks while our concept signature is used for image annotation and semantic image retrieval.

2.3 Finding Communities in Networks

Network structure has drawn great attention in analyzing social relationships between people. Network structures are proposed in [36, 188] as interaction graph where individuals are indicated by the nodes and the edges between them are weighted by their relatedness in either social or visual sense. A different type of network is presented in [23] with edges express the probability of individuals belonging to a group.

A very common property in many realistic complex networks such as social networks and biological networks is known as the *community structure* [69, 122], i.e., the nodes in the network naturally divide into groups with denser connections inside each group and looser connections among groups. In our tracklet interaction network, the nodes and edges represent pedestrian tracklets and their social grouping behavior, respectively, and the social groups can be viewed as the communities in the network.

Traditional algorithms for detecting groups of nodes in a network can be categorized into partition based methods [59], hierarchical clustering algorithms which can be further classified into agglomerative (e.g., [12]) and divisive (e.g., [116]) algorithms, spectral algorithms [113], modularity-based methods [118], and dynamic algorithms [77]. In most of the work [12, 59, 118] the edges are unweighted in the problem domains, thus, additional computing is required, e.g., in [46] the edge weight is defined by the number of non-independent paths between nodes which can be computed using polynomial-time "max-flow" algorithms, and in [69] it is defined by Freeman's edge betweenness centrality. However, in our network the edge weights are computed directly from the distance metric defined on the spatio-temporal relations between tracklets. As a single node (tracklet of pedestrian) can be present in multiple groups simultaneously (the uncertainty of social groups, for example, a tracklet has equal distances to the other two tracklets), this results in the overlapping of groups, or the sharing of nodes between groups. There are techniques devoted to solve this problem in recent network analysis research [123, 194].

Nodes and edge weights can change over time when the video sequence proceeds. The emergence of new groups as well as the growth, split, merge, and death of old groups can occur over time. As compared to the other algorithms, modularity based approaches have been demonstrated to be the most effective in finding good partitions in an efficient manner in large networks, and they can address weighting, overlapping and evolving problems in a network [68], therefore, we adopt modularity-based approach in our work to find the social groups from tracklet interaction networks.

2.4 Automated Insect Identification/Retrieval Systems

Insect species identification recently has received great attention due to the urgent need for systems that can help in biodiversity monitoring [130], agriculture and border control [6, 93], as well as conservation and other related research [154]. Likewise, identifying species is also the prerequisite to conducting more advanced biological research such as species evolution and developmental studies. However, the vast number of insect species and specimen images is a challenge for manual insect identification. The request for automated computer systems is only likely to grow in the future.

Several attempts have been made in the last two decades to design species identification systems from any type of available data. There have been sophisticated applications to solve problems in classifying orchard insects [176], recognizing the species-specific patterns on insect wings [60] and identification of insect morphologies on fossil images [81], etc. It has been recognized that these computer-aided systems can overcome the manual processing time and errors caused by human subjectiveness.

Besides the above mentioned systems, there are more well-known systems: the SPecies IDentification Automated (SPIDA) system [39], the Digital Automated Identification SYstem (DAISY) [121], the Automated Bee Identification System (ABIS) [144] and DrawWing [157], a program for describing insect wings in a digital way. The first two systems use machine learning techniques such as neural network as the core of the classifier, while DAISY is not only used for moth identification as the main purpose but also used for any type of species identification in general, such as fish, pollen and plants. On the other hand, SPIDA is designed for recognizing 121 spider species in Australia. The system keeps refining its learning accuracy by using user uploaded labeled images as more training data. ABIS uses a similar idea as us on finding attribute patterns from bee's wings to recognize their species. It utilizes the SVM-based discriminative classifier and the average performance reaches 95% in accuracy.

One common property of these systems is that they all rely on images taken from carefully positioned target under consistent lighting conditions which reduces the difficulty of the task to some extent. One interesting aspect of automated species identification is that the data are not limited to images. For example, the paper proposed by Ganchev et al. [63] describes the acoustic monitoring of singing insects that applies sound recognition technologies into the insect identification task. Meulemeester et al. [110] report on the recognition of bumble bee species based on statistic analysis of the chemical scent extracted from the cephalic secretions. A challenging competition on multimedia life species identification [84] was recently held on identifying plant, bird and fish species using image, audio and video data separately.

The development of these systems made great efforts in incorporating machine learning techniques like principal component analysis (PCA), linear discriminant analysis (LDA), artificial neural networks (ANNs), support vector machine (SVM) and many other techniques.

With the increase of insect images, there is a growing tendency in the field of entomology by using image retrieval systems to help archive, organize and find images in an efficient manner. Content-based image retrieval [151] has been well studied and developed for many years in the image retrieval domain. It looks at the contents of the image itself and extracts certain pictorial features used to compare the image similarity automatically. Great efforts have been made using content-based image retrieval technique to find the relevant images to a query based on the visual similarity, the prototype systems for retrieving Lepidoptera images include "butterfly family retrieval" [168], a web-based system "Butterfly Ecology" [169] and a part based system [11]. Most of these systems focus on extracting low-level features such as color, shape and texture as the image representation that allows the systems to compare images based on these features.

These systems are attractive but still present a number of problems. For example, a powerful function of CBIR is the ability to integrate user interaction where retrieval precision is adjusted according to user provided relevance feedback (RF) information. However, none of the existing systems has adopted the RF scheme into the retrieval framework. Also, a common limitation of the available systems is they only cope with a comparatively small number of species or categories in the dataset.

2.5 Multi-Pedestrian Tracking with Social Groups

In Chapter 5, we apply our concept co-occurrence pattern detection model in a very special application senario which is known as multi-pedestrian tracking with social groups. People tend to form groups when they walk. If we consider indivisuals as nodes and their social relations during walk as edges, we can actually represent pedestrians in a network and the social groups among people can be viewed as the co-occurrence patterns in the network structure. We propose to understand the social grouping behavior based on current computer vision techniques for pedestrian detection, multi-people tracking and data association to concatenate short tracks into longer reliable trajectories passing through the scene. State-of-the-art multi-people tracking approaches can be categorized into two classes based on the time sensitivity: real-time tracking and time-delayed tracking. In real-time tracking, the detection responses and the correspondences among them are usually jointly estimated and updated for each frame by using the information acquired from previous frame. Techniques such as particle filter are often adopted [8, 16] to estimate the intermediate states. Many approaches in this category focus on tracking each target separately [170] and they tend to fail when encountered with challenging situations involving from inter-people and scene occlusions, illumination or appearance variations and abrupt motion changes. However, there are also approaches such as [8] that jointly track individuals and groups and demonstrate that individual tracking can be improved by group tracking and *viceversa*.

For the approaches in time-delayed tracking category, multiple targets are tracked simultaneously [4, 153]. The detection responses produced by pedestrian detectors are formed into tracklets and the final tracks are obtained by associating the tracklets at different granularities [78]. The association of tracklets is addressed by global optimization solutions such as K-shortest path [9], Hungarian algorithm [78], CRF [180] and cost-flow network [193]. The occlusions are modeled as merging and splitting of tracklets and solved by using Markov Chain Monte Carlo (MCMC) [187]. Most of these approaches generally do not use high-level semantics such as social groups to improve data-association for tracking.

Discovering the interactions among pedestrians to improve tracking in crowded scenes has become a new trend of research in the literature. Solmaz et al. [152] introduced an approach that identifies individual/group behaviors without any object detection, tracking or training steps. Pelligrini et al. [129] proposed a dynamic model for tracking people in complex scenes that exploit the social interactions such as attraction and repulsion. According to recent research by Moussa *et al.* [114], 70% of people in a crowd walk in groups. The grouping property of pedestrians is explicitly analyzed in the computer vision field in [65]. Specifically, groups are used as contextual knowledge for trajectory prediction and refinement [128, 179].

With the increasing need for surveillance systems monitoring and detecting activities of interests in mass events with their continuing growth in size and frequency, the study of social grouping behavior of pedestrians by using computer vision techniques has become a popular research area [22, 88, 129, 109, 172, 191]. When people walk, they naturally form groups with smaller distances to the members in the same group and larger distances to the people outside the group [114]. An interesting discovery by MacPhail shows that 89% of people attend events in groups and 94% of them leave with the people they come with [107].

Members in the same group often share same walking behaviorals, known as the *collective behavior* of pedestrians [17], such as change of direction and speed, way of avoiding obstacle, etc., that describes the distinctive and dramatic features of group trajectories and of individual trajectories within groups. In turn, groups can be determined based on the individual spatial location, cardinality and velocity [67]. Recent research efforts [28, 29, 196] have suggested that social groups that exhibit collective behaviors can be used to improve the understanding of social events in video sequences involving interactions among groups, especially in the cases where the cameras have elevated viewports and monitor crowded environments in which pedestrians are still discernible while partial body occlusions happen frequently. In the context of role understanding of social groups in video sequences, many approaches [25, 36, 37, 55, 137, 189] have been proposed that combine sociological analysis and computer vision techniques to detect and recognize the behaviors of social groups by using key frames extracted from a video.

There is recent evidence that more efficient algorithms can be developed based on the recognition of high-level social groups detected in a hierarchical structure [142, 195]. The social grouping behavior of people shopping together is captured and evaluated by analyzing the inter-body distances [75]. The velocity similarity has been applied in [135, 179] to group people together for motion prediction and tracking. Ge *et al.* [64] identify small groups of pedestrians based on pre-detected trajectories, however, unlike our approach, they model the social grouping behavior in a pairwise manner, and they overlook the dynamic structural changes of the social groups (merge, split, appear, disappear, etc. [175]).

Chapter 3

Semantic Concept Co-occurrence Patterns for Image Annotation and Retrieval

3.1 Introduction

Semantic concept inference addresses the problem of deriving concepts from multimedia visual content. It is an essential ingredient for many applications, such as automated image annotation and concept-based image retrieval. The concepts cover a variety of topics, such as a single object (e.g., *table*), a scene (e.g., *beach*), an event (e.g., *commencement*), and a piece of knowledge (e.g., drive car). The difficulty of detecting these concepts varies as the semantic complexities are different.

Approaches for detecting semantic concepts attempt to address the fundamental issue of bridging the semantic gap [104]. Recent research has demonstrated the effectiveness of using semantic correlation instead of visual correlation as a contextual cue to narrow the semantic gap in the multi-concept inference task. For example, individual concept detector can be confused by the concepts having very similar visual properties, e.g., “ocean” and “sky”. The explicit knowledge from contextual information, e.g., “sky” and “airplane” have a stronger correlation than “ocean” and “airplane”, can actually help reduce or even remove the uncertainty in the inference results.

In general there are five types of correlations between semantic concepts: **(i)** Synonymy, **(ii)** Similarity, **(iii)** Meronymy **(v)** Inclusion and **(iv)** Co-occurrence. The first four types measure the relationship between the semantic meanings of the concepts, while co-occurrence measures the concurrent frequency of concepts.

Compared to other four types of correlation that have been widely used in the concept inference literature, co-occurrence can be more important in the inference of complex scenes because multiple concepts are cocurrently presented. For example, the uncovered patterns of co-occurrences can be used to help distinguish visually similar objects based on the context, e.g., horse could be reinforced by the pre-discovered co-occurrence pattern (horse, windmill), and zebra could be weakened because it has no co-occurrence relationship with windmill. The underling idea is that concepts that are observed together across many images are likely to have the co-occurrence relationship and constitutes a co-occurrence pattern. For example, if a group of traffic light, street, car is observed in many images, we consider it as a co-occurrence pattern, and the inference of each individual concept can be improved if we have the knowledge of the co-occurrence patterns in advance.

The problem is how can we discover the co-occurrence patterns that realize the underlying groud-truth in maximum accuracy and efficiency. Considering the large number of semantic concepts in the real world and their intricate co-occurrence relationships, it would be natural to represent them in a network with nodes indicating the concepts and edges standing for the interactions between nodes. The edges can be naturally associated with weights to denote the significances of the co-occurrences. A common property that has been discovered in many networks is the *community structure* property, which is the partition of network nodes into groups (communities) with highly inter-connected nodes (more edges with larger weights), with nodes belonging to different groups being sparsely connected (less edges with smaller weights). Inspired by the theories in network analysis, discovering the concept co-occurrence patterns, which are the groups of concepts that co-occur frequently, can be solved by identifying the communities, which are sets of closely connected nodes in the network representation.

We qualitatively define the *communities of concepts* in the network as groups of nodes (concepts) that have tight internal connections (co-occurrences) and loose external connections (co-occurrences) to the other groups. Therefore, a *hierarchical community structure* can naturally reflect the co-occurrence patterns at different semantic levels. We illustrate the idea of hierarchical community structure and co-occurrence pattern in Figure 1.1.

Graph partition algorithms provide an effective alternative for analyzing the communities. We adopt modularity optimization [118] in our framework to uncover the communities. The final goal is to utilize the detected hierachical co-occurrence patterns (communities)to boost the accuracy of individual concept inference in different applications such as automatic image annotation and concept-based image retrieval. We demonstrate the effectiveness of our approach on a wide variety of concepts in real images obtained from popular benchmark datasets. Experimental results in the proposed application scenarios show clear gains from co-occurrence patterns comparing to other baseline approaches with/without exploiting concept correlations.

Figure 3.1 shows the system diagram of the proposed approach.

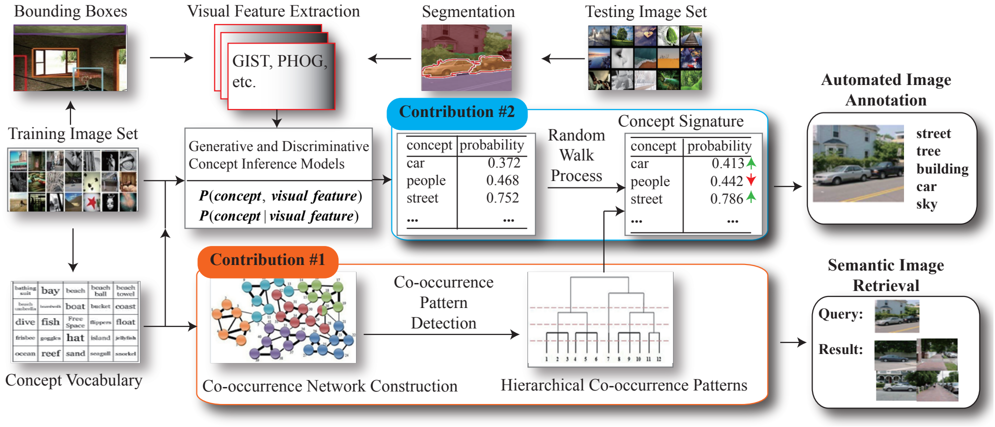


Figure 3.1: The flowchart of the proposed concept inference framework. The contributions are: (i) a co-occurrence pattern detection method that effectively explores hierarchical correlations among semantic concepts, (ii) random walk based approach to refine the concept signature representation based on detected concept co-occurrence patterns.

3.1.1 Contributions of This Chapter

In constrast with state-of-the-art approaches, we summarize the fundamental **contributions** of this chapter below:

1. We devised an original approach to discover the hierarchical co-occurrence patterns of concepts as underlying community structures in a concept co-occurrence network. This is the first work as we know that attempts to explore the co-occurrences from the network analysis point of view. The co-occurrence patterns or communities capture the concept concurrent property and provide more information for individual concept detection. Accordingly, we propose method to utilize the detected patterns to improve individual concept detection, infer and boost more difficult concepts that usually have large visual variations from easy ones in complex scenes.

2. We introduce a simple image content descriptor referred to as *concept signature* generated from the concept detection responses. Neither like traditional visual

descriptors that are purely based on low-level pictorial features, nor like textual descriptors such as labels, captions, keywords that contain only the high-level semantic information, the proposed descriptor can record the semantic concept with its corresponding confidence value inferred from low-level features. Based on the concept signature, we can annotate images with the concepts that exceed certain confidence threshold. We can also estimate the semantic distance between two images under a given metric.

To leverage the contextual information, we only deal with the images with multiple concepts. In order to acquire a reliable individual concept detector, the training images are labeled at the object level, i.e., the concepts are given with the minimum bounding rectangles (MBRs), and the visual features are extracted regionally. In the semantic sense, a pool of concepts is collected from the training set as the vocabulary to construct the co-occurrence network (described in Section 3.2.1.1) for concept co-occurrence pattern detection (Section 3.2.1.2).

The semantic concepts are used to build probabilistic models for inferring the correspondence between a semantic concept and the relevant visual features (Section 3.2.2). We use both generative and discriminative models, for comparison, as individual concept detectors to discover the semantic concepts in the test images.

Concept signature is introduced as visual and semantic description of images with its elements obtained from the individual concept inference results (described in Section 3.2.3). With the help of the uncovered concept co-occurrence patterns, the concept signature is further refined to approach the ground-truth labels through a random walk process. The effectiveness of the proposed framework is evaluated experimentally in Section 3.3 for automatic image annotation and concept-based image retrieval applications. Table 3.1 summarizes the definition of symbols used in Section 3.2.

3.2 Technical Details

3.2.1 Construction of Co-occurrence Network and Pattern Detection

3.2.1.1 Co-occurrence Network Construction

In this section, we discuss the representation of various co-occurrence relationships among different semantic concepts. As the number of concepts is large and the relationship among them tend to be complex, we model them by a network structure. In this paper, we name such a network of concepts as *Concept Co-occurrence Network* (CCN). Let $G = (V, \omega)$ represent a network structure, where each edge $e \in E$ is assigned with a positive weight $\omega(e)$ corresponding to its importance in the network. Let $\Phi = \{c_1, c_2, ..., c_m\}$ be the concept vocabulary in the training image set, where m is the total number of unique concepts annotated to the images that the system is attempting to detect. Let $T = \{t_1, t_2, ..., t_n\}$ denote the training image set with size n. The CCN is constructed by associating each concept c_i with a node v_i in G. Concepts with textual and visual appearances in the same media resource are likely to have co-occurred and should be linked together by an edge in E.

The edge weight is determined by three types of co-occurrence measure, namely, *global semantic co-occurrence* measures, *global visual co-occurrence* measure and *local visual co-occurrence* measure. First, We evaluate the global semantic co-occurrence by the normalized Google distance [30] (NGD). NGD is proposed to compute the pairwise conceptual distance by counting the number of web pages containing the query concept returned by Google search engine. NGD is intrinsically a co-occurrence measure that explores the co-occurrence of words from on-line textual documents assuming a global

Table 3.1: Definition of symbols used in Section 3.1

Symbols	Definitions
$G(V,\omega)$	The constructed concept co-occurrence network with V and E representing the node and edge sets separately, and ω denoting the edge weight.
v_i	The i^{th} element in the node set V.
Φ	The vocabulary of semantic concepts in this work.
m	The number of semantic concepts in vocabulary Φ.
c_i	The i^{th} element in the concept vocabulary Φ.
T, t_i	The training image set and the i^{th} element.
n	The number of images in the training set.
$A_{m\times m}$	The adjacency matrix used to record the edge weights in G, $A(c_i,c_j)$ denotes the weight of the edge connecting concepts c_i and c_j.
$H_{m\times n}$	The association matrix, $h_{ik}=1$ if concept c_i appears in image t_k in the training set and 0 otherwise.
G(c)	The number of pages containing concept c reported by Google search engine.
G(c_1, c_2)	The number of pages containing concepts c_1 and c_2.
Ω	The number of pages indexed by Google.
F(c)	The number of images containing concept c in Flickr.
F(c_1, c_2)	The number of images containing both concepts c_1 and c_2 in Flickr.
Ψ	The number of images indexed by Flickr.
$x_{i,k}$	Equals 1 if concept c_i appears in training image t_k, 0 otherwise.
η_1,η_2,η_3	The weights set to evaluate the importance of each co-occurrence measure, $\sum_{i=1}^{3}\eta_i=1$.
C	The community detected in the network structure.
Q_C	The modularity measure of community C.
ΔQ	The modularity gain acquired when the community structure changes.
k_i	The summation of edge weights attached to node v_i in the network.
$k_{i,C}$	The summation of edge weights where the edges are connecting node i to the nodes in community C.
Γ	The half of the summation of all the edge weights.
δ	The delta function used in computing the modularity.
Σ_{in}	The summation of edge weights inside community C.
Σ_{out}	The summation of edge weights that link to the nodes outside community C.
Λ_c	The visual variation of semantic concept c.
R_c, $\lvert R_c\rvert$, r_c^i	Training region set containing concept c, the size of the set and the i^{th} element.
$R_{\bar{c}}$, $\lvert R_{\bar{c}}\rvert$, $r_{\bar{c}}^j$	Negative training region set of c, the size of the negative set and the j^{th} element.
f_{R_c}	The mean of the feature vectors of the regions in R_c.
$f_{r_c^i}$	The feature vector of i^{th} region r_c^i in R_c.
$f_{r_{\bar{c}}^j}$	The feature vector of j^{th} region $r_{\bar{c}}^j$in $R_{\bar{c}}$.
Z	The dimension of the above feature vectors.
D_{χ^2}	The Chi-square distance between two feature vectors.
$\mathcal{G}$	The function that generates the prototype vector.
g	The prototype vector generated from a region.
w	The weight vector in the SVM objective function.
b	The bias vector in the SVM objective function.
e_1,e_2	The constants for controlling the relative influence of the two competing terms in the SVM function.
h	The hinge loss function in the SVM objective function.
CS	The concept signature descriptor.
s_{c_i}	The confidence score of concept c_i in the signature.

Table 3.2: The summarization of the usage and motivation for adopted co-occurrence measures.

Co-occurrence Measure	Usage & Motivation
Normalized Google Distance (NGD)	Captures the global semantic co-occurrences. The number of semantic concept co-occurrences in a local dataset is far below than what is generated by the massive web users. For example, there are 434 million concepts/annotations found from web images [35]. NGD can actually reflect the confidence that two semantic concepts can co-occur among online textual resources.
Normalized Tag Distance (NTD)	Captures the global visual co-occurrences. NGD assumes concept relationships only depend on semantic co-occurrences in the text field which cannot guarantee the existence of these co-occurrences from the visual perspective (i.e., the presence in the images). NTD treats the tags that are associated with the images as the general semantic concepts used in NGD and calculates the co-occurrence in the same way as NGD. It strengthens the visual co-occurrences between concepts.
Automatic Local Analysis (ALA)	Captures the local visual co-occurrences. NGD and NTD utilize the global information that is out of the scope of a local dataset. However, global co-occurrences may not exactly match the local co-occurrence in an image collection. Therefore, ALA is introduced to strengthen the local visual co-occurrences.

meaning of words. Second, for global visual co-occurrence measure, we adopt Flickr based normalized tag distance [100] (NTD) measure. NTD treats the tag list associated with each image in a role similar to the web page in NGD and it calculates the conceptual distance in the same way. Since tag lists indicate visual co-occurrences of concepts in social media resources, it is very intuitive to use NTD to reflect the global frequency of concept co-occurrences. Finally, we apply automatic local analysis [7] (ALA) to identify local visual co-occurrence of concepts in a particular image dataset denoted as the training set in order to capture the local co-occurrence property in the specified image collection. The motivations and the usefulness of the three measures are summarized in Table 3.2.

The motivation for using the three co-occurrence measures is that they can complement one other. NGD uses the entire World-Wide-Web as the dataset which is known to be the largest on earth. The contextual information is given by billions of independent persons of knowledge, thus, it can overcome the limitation in the scope of the

```
Algorithm 1: CCN construction
Input: Training image set $T$ with $n$ images, a vocabulary $\Phi$ with $m$ individual concepts
Output: Constructed concept co-occurrence network $G = (V, \omega)$
Initialize a $m \times m$ concept adjacency matrix $A$ for recording edge weights with every element set to 0.;
Measure the global semantic co-occurrence between each pair of concepts $\{c_i,\ c_j\}$, $i \in 1, ..., m$, $j \neq i$ by normalized Google distance [30]: $NGD(c_i, c_j) = \frac{max\{logG(c_i),\ logG(c_j)\} - logG(c_i,\ c_j)}{log\Omega - min\{logG(c_i),\ logG(c_j)\}}$;
Measure the global visual co-occurrence by normalized Tag distance [100]: $NTD(c_i, c_j) = exp\frac{max\{logF(c_i),\ logF(c_j)\} - logF(c_i,\ c_j)}{log\Psi - min\{logF(c_i),\ logF(c_j)\}}$;
Measure the local visual co-occurrence by automatic local analysis [7]: $ALA(c_i, c_j) = exp(-\Delta)$, where $\Delta = \frac{\sum_{t_k \in T} x_{ik} \times x_{jk}}{\sum_{t_k \in T} x_{ik} \times x_{ik} + \sum_{t_k \in T} x_{jk} \times x_{jk} - \sum_{t_k \in T} x_{ik} \times x_{jk}}$;
Combine the three measures into the final co-occurrence significance and assign the value to element $A(c_i, c_j)$;
$A(c_i, c_j) = \eta_1 \cdot NGD(c_i, c_j) + \eta_2 \cdot NTD(c_i, c_j) + \eta_3 \cdot ALA(c_i, c_j)$. In our setting we put equal importance on the three measurements, so $\eta_i = \frac{1}{3}$;
Traverse all the elements in $A$, add $c_i$ as node, connect two nodes $c_i, c_j$ with edge weight according to the value of $A_{ij}$;
```

concepts represented in image datasets. However, NGD does not involve any visual information in the distance calculation, and co-occurred concepts in the textual documents may have zero probability to appear in the real life images (e.g., concepts from *science-fiction novels*). Therefore, visual co-occurrences are analyzed to decrease the ambiguities arisen from texts. Global visual co-occurrences from community-contributed web image collections, e.g., Flickr, are represented by the rich tags as metadata. However, it cannot accommodate the changes to the training dataset. i.e., images and concepts that are added or removed from the original dataset. Local visual co-occurrence can contribute to dynamic dataset, thus, it is reasonable to be considered. The steps for constructing the CCN are described in Algorithm 1.

3.2.1.2 Co-occurrence Pattern Detection

Finding the co-occurrence patterns of the interconnected nodes corresponds to uncovering community structures from the randomness of the network topology which is close to graph clustering or partitioning problem. However the problem is computationally intractable. Recently *modularity* has been used as a criterion for determining the effectiveness of the detected communities, and at the same time it can serve as an objective function to maximize. In this paper we adopt *modularity optimization* paradigm to address the problem and propose a method based on Newman-Girvan modularity [118] optimization. The modularity measures the quality of a partition by comparing the link density of nodes inside a community with the links to the outside nodes. Usually high values of modularity suggests good partitions. In the case of weighted network, we define the modularity of community C as:

$$Q_C = \frac{1}{2\Gamma}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2\Gamma}]\delta(ID_i, ID_j) \tag{3.1}$$

Typically modularity score is in the range of $[-1, 1]$, and in practice a value greater than 0.3 indicates a significant community. The modularity is calculated over all the pairs of nodes in the network, where ID_i and ID_j are their community IDs, $\delta(ID_i, ID_j) = 1$ if $ID_i = ID_j$ for two nodes v_i and v_j, otherwise $= 0$. We consider iteratively merging the nodes into a hierarchical community structure with different levels of resolution by maximizing the modularity gain at each iteration. The *modularity gain* of moving an outside node v_i into a community C is evaluated by:

$$\begin{aligned}\Delta Q = [&\frac{\Sigma_{in} + k_{i,C}}{2\Gamma} - (\frac{\Sigma_{out} + k_i}{2\Gamma})^2] - \\ &[\frac{\Sigma_{in}}{2\Gamma} - (\frac{\Sigma_{out}}{2\Gamma})^2 - (\frac{k_i}{2\Gamma})^2]\end{aligned} \tag{3.2}$$

Please see Table 3.1 for the definitions of symbols. Algorithm 2 is given for detecting the hierarchical concept co-occurrence patterns (communities) in a network. The runtime of the algorithm for co-occurrence pattern detection is $O(|V|(|E| + |V|))$ where $|E|$ is the number of edges and $|V|$ is the number of nodes in the network. The algorithm iteratively generates a hierarchical community structure with different resolutions, in other words, the communities of individual concepts, and the communities of communities. To point out, our algorithm addresses the share of nodes problem between communities explicitly.

3.2.2 Concept Occurrence Inference Models

We integrate the detected concept co-occurrence patterns for individual concept inference. We use probabilistic inference models to build the correspondence between concepts and regional visual features from training data. The outputs of the model running on testing images are vectors of concepts with corresponding probabilities scores of the occurrence. We name this vector representation as *concept signature* which captures both the semantic and visual information about images.

Individual concept inference is the baseline and key factor to the overall performance although we demonstrate later that it can be improved by utilizing the co-occurrence patterns. In order to compare the effect of the baselines, we implement two individual concept inference models based on generative and discriminative training.

The ***generative model*** is built by jointly estimating the probability of visual and semantic representations. Suppose T is the training set of annotated images and R is the set of corresponding segmented regions, and let r be an element of R. We specify the process of feature vector generation and vector quantization as an integrated function $\mathcal{G}$, with $g = \mathcal{G}(r) \in \mathbb{R}^J$. Each image t in T can be represented as a set of regions $r_t = \{r_1, r_2, ..., r_n\}$ along with the corresponding concept from the set $\{c_1, c_2, ...c_n\}$.

Algorithm 2: Concept co-occurrence pattern detection

```
Input: Co-occurrence network built from Algorithm 1
Output: Hierarchical concept co-occurrence patterns
while Positive Modularity Gain can be achieved do
    Partitioning phase;
    foreach $c_i$ in the vocabulary($i = 1, ..., N$) do
        Assign Node $n_i$ represents $c_i$ in the network;
        Label $n_i$ with community tag $C_i$;
    Each node will have an unique community tag after above step;
    while Positive Modularity Gain can be achieved do
        foreach $n_i$ in the network do
            Remove $n_i$ from its original community $C_i$;
            foreach neighboring community $C_j$ of $n_i$ do
                Add $n_i$ to $C_j$;
                Calculate modularity gain $\Delta Q$ (eq.(2)) after changing the community structure;
                if $\Delta Q > 0$ then
                    Let $C_{old}$ and $C_{new}$ denote the original community and new community of node $n_i$;
                    Compute modularity scores $Q_{C_{old}}$ and $Q_{C_{new}}$ by eq.(1);
                    if $Q_{C_{old}} >= 0.3$ and $Q_{C_{new}} >= 0.3$ then
                        $n_i$ is shared by both communities;
                        Split $n_i$ into $n_i$ and $n_i^{'}$;
                        Add $n_i$ into $C_{old}$ and add $n_i^{'}$ into $C_{new}$;
                        Copy the edges of $n_i$ that are incident to other nodes for $n_i^{'}$;
                    else if $Q_{C_{old}} < 0.3$ and $Q_{C_{new}} >= 0.3$ then
                        Change the community tag of $n_i$ from $C_{old}$ to $C_{new}$;
                    else
                        $n_i$ stays in the original community;
                else
                    $n_i$ stays in the original community;
    Coarsening phase: generates the hierarchical structure;
    Replace the nodes in the same community detected from the above steps as a single node;
    Replace the edges between the nodes in two adjacent communities by a single edge with summed edge weights;
    Represent edges in the same community as a self-looped edge with weight equal to the sum of the internal edge weights;
```

Given an image region r, *firstly*, we model the probability of obtaining concept c by sampling from a multinomial distribution $P_{\mathcal{M}}(c|r)$ that will split probability mass among multiple concepts. Subscript $\mathcal{M}$ represents the multinomial distribution. *Secondly*, we model the relation between a region r in the training set and a possible prototype vector g as a distribution $P_{\mathcal{R}}(r|g)$. *Finally*, when given a region r from the unknown set, we model the probability of getting a prototype vector g by sampling from a distribution $P_{\mathcal{G}}(g|r)$. For an unknown region r_i from a test image, the probability of observing c_j is given by the joint probability:

$$\begin{aligned} P(r_i, c_j) = \sum_{r_t \in R_{c_j}} & \Big\{ P(r_t) \cdot P_{\mathcal{M}}(c_j|r_t) \\ & \cdot \Big\{ \sum_{g_t} P_{\mathcal{R}}(r_i|g_t) \cdot P_{\mathcal{G}}(g_t|r_t) \Big\} \Big\} \end{aligned} \tag{3.3}$$

We assume that the training set is sufficient to cover all possible instances of the region-concept pair in the test set. The larger the size of the training set, the more correct knowledge about the generative model that we can obtain.

The ***discriminative model*** is created by an ensemble of instance-SVMs for each concept where the idea is similar to [105]. For each concept, the positive instances are the regions containing that concept and the rest are negatives. We first train a separate linear SVM classifier for each positive instance of a given concept with the negatives. For each positive instance with feature $f_{r_c^i}$ of concept c, and the negative set $R_{\bar{c}}$ with instance feature $f_{r_{\bar{c}}^j}$, the weight vector w are learned by optimizing the convex objective:

$$\mho(w, f_{r_c^i}, b) = ||w||^2 + e_1 h(w^T f_{r_c^i} + b) + e_2 \sum_j h(-w^T f_{r_{\bar{c}}^j} - b) \tag{3.4}$$

where h represents the hinge loss function $h(x) = (0, 1 - x)$ which permits hard-negative mining to find the small subset of dominating negative support vectors from $R_{\bar{c}}$. For a test region, the instance-SVM classifiers of a concept are first applied. The outputs from individual classifiers are fused by weighted averaging to generate the final concept score. The weight w attached to each single classifier is determined by adaptive linear neural network (ALNN) in a validation process.

3.2.3 Concept Signature and its Refinement

We propose concept signature as image descriptor. Concept signature is a vector in which each entry contains a tuple of concept and its occurring probability from the inference model. Compared to other image descriptors, concept signature: 1) records both the visual and semantic information of an image, thus, image can be compared and retrieved based on high-level semantic concept similarity, which we denote as *concept-based image retrieval* in this paper. 2) has a very simple form, therefore, it can lower the memory cost for storing large image collections and decrease the computational costs. 3) can keep all the concept occurrence probabilities which can be revised later on when individual concept inference accuracy is improved.

We refine the original scores in the concept signature in a *re-ranking* manner formulated as a random walk process over the contextual co-occurrence patterns. Suppose the hierarchy has L levels, we set the lowest level that contains the semantic concepts as level-1 and the highest level as level-L. Assume initially concept c_i has occurring score s_{c_i} given by the inference model, and let $lowest(c_i, c_j)$ denote the function to compute the level of the lowest superordinate (common ancestor) between c_i and c_j. In the k^{th} updating iteration, the score s_{c_i} is refined by the random walk process:

$$s_{c_i}^k = \alpha \sum_{c_j \neq c_i} s_{c_j}^{k-1} \cdot \frac{lowest(c_i, c_j)}{L} + (1 - \alpha) \cdot s_{c_i}^{k-1} \tag{3.5}$$

We set α to 0.5 which means the effects from its own score and the scores from neighboring concepts are treated equally. The scores are updated recursively until all the scores converge. Eq. 3.5 can strengthen the scores of concepts in more closely related patterns and weaken the more isolated ones. Finally, we give Algorithm 3 for generating image concept signature and random-walk refinement:

Algorithm 3: Concept signature refinement

```
Input: Testing image set
Output: Refined concept signature representation for each testing image
foreach Image $T$ in the testing set do
    Detect the salient regions $r_1, ..., r_m$ by mean shift based segmentation [31];
    foreach Salient region $r_i$ do
        Apply the inference models defined in eq. 3.3 or eq. 3.4;
        Compute the original regional signature $CS_{r_i} = ((c_1, s_{c_1}), ..., (c_n, s_{c_n}))$;
    Compute the intermediate image-level signature by $CS_I = \frac{1}{m} \sum_{i=1}^{m} CSr_i$;
    Obtain the final image concept signature by random walk based refinement (eq. 3.5);
```

3.3 Experimental Results

3.3.1 Image Datasets and System Parameters

3.3.1.1 Image Datasets

• The **LabelMe** [140] dataset is a collection of 72,852 images containing more than 10,000 concepts. We use a subset which contains 10,000 images and 2,500 concepts. The raw images have different resolutions (e.g. 2560 × 1920, 1600 × 1200, 256 × 256, etc.). In this paper, we use the resolution of 1600 × 1200 downloaded from the website by using the Toolbox provided by the dataset creators.

• The **Scene Understanding (SUN'09)** [27] dataset contains 12,000 images and more than 5,800 concepts covering a variety of indoor/outdoor scenes. The total number of annotated labels is 85,456 which results in an average of seven labels per image. The images are collected from multiple sources (Google, Flickr, Altavista, LabelMe) and are labeled by a single annotator using the LabelMe tool. The labels are manually verified for consistency.

• The **Outdoor Scene Recognition (OSR)** [120] dataset has 2,682 images with 520 concepts across eight outdoor scene categories: coast, forest, highway, inside-city, mountain, open-country, street, tall-building. All the concepts are labeled with corresponding bounding boxes manually.

The selected datasets have the following advantages compared to other datasets (e.g., TinyImages [159], MSRC [146], Caltech-101 [94]): **(i)** All the datasets present complex scenes containing multiple concepts in a single image which is suitable for exploring the concept co-occurrence correlations. **(ii)** Compared to the general and specific terms defined in the *synonym set* in WordNet (e.g., "mammal", "tool", "geological formation") and used by ImageNet (e.g., "coconut tree", "ocean floor", "Davy Jones"), most of the concepts are at the intermediate level of semantics (e.g., "tree", "sea", "people") which are more relevant to Folksonomy-style tags used in daily life. **(iii)** The datasets have a large number of concepts that cover a great majority of object categories. **(iv)** The bounding boxes for the concepts are available in standard XML format which can be easily parsed by programs (e.g., the open source tool TinyXML used in our framework).

3.3.1.2 System Parameters

The weighting parameter ω (Section 3.2.1.1) is set to 1/3 for the three measures. The modularity threshold Q_C (Section 3.2.1.2) is set to 0.3, and the weight parameter α in the random walk process (Eq. 3.5) in this paper is set to 0.5. All the parameters are set empirically and they are kept constant for all the experiments reported in this paper.

3.3.2 Visual Features

We extract visual features locally from the regions enclosing the concepts defined by minimum bounding rectangle (MBR). For test images, the features are extracted from the MBRs of the segmented salient regions. The features are:

- **Color GIST** feature [120] is computed on 4×4 grids over the concept bounding box. The MBRs are resized to 32×32 (we do not maintain the aspect ratio) and then the orientation histograms are calculated at 3 scales with 8, 8 and 4 bins.

- The **pyramid of histogram of oriented gradients (PHOG)** feature [13] is computed by following steps: 1) extract the Canny edges in the concept bounding box, 2) quantize the gradient orientation on the Canny edges (from 0° to 180°) into 20 bins, 3) Four spatial pyramid levels are used (1×1, 2×2, 4×4, 8×8). Each level is used in an independent kernel.

- **PHOG with oriented edges** [161] considers the direction (0° to 360° divided into 40 bins) of the salient Canny edges. We use four-level spatial pyramid.

- The **pyramid of Shechtman and Irani self similarity** feature [145] is computed at every 5 pixels and quantized into 300 clusters using k-means, and then the histograms are calculated at three levels.

- The **bag of visual words** feature [161] is obtained by first computing the SIFT descriptors [102] at the interest points detected by Hessian-Affine detector [89], and then quantizing them into a vocabulary of visual words with the size of 1000. Finally, a sparse histogram is generated based on the visual words.

3.3.3 Applications and Evaluation Criteria

3.3.3.1 Application 1: Automatic image annotation

The goal is to predict concept occurrences for an image from a given concept vocabulary. The predictions are then used to annotate the image based on the rank of the probability scores. Most existing approaches for AIA neglect the co-occurrence patterns among concepts and annotate the concepts individually. In our framework, the concepts ranked as top-M in the refined concept signature based on the inferred probability scores are used as the annotations. An alternative way with unfixed annotation length is to use all the annotations with scores passing certain threshold.

3.3.3.2 Application 2: Concept-based image retrieval

For a given query, we compute the similarity to the database images based on the concept signature representation using the Earth Mover's Distance (EMD) [139] as the distance metric. Given two concept signatures $\mathbf{p}$ and $\mathbf{q}$, the EMD is defined as: $\mathrm{EMD}(\mathbf{p},\ \mathbf{q}) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} o_{ij} d(p_i, q_j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} o_{ij}}$, where o_{ij} denotes the flow and it follows the constraints of the scores in the concept signature and $d(p_i, q_j)$ is the pre-defined ground distance between each pair of individual concepts. In our setting, we use the reciprocal of the edge weight in the co-occurrence network as the measure of ground distance. EMD measures the least amount of work to completely transfer one signature into another, it is calculated by linear programming [139].

3.3.3.3 Evaluation criteria

- Automatic image annotation: The performance is evaluated by *Top-M* $F_{0.5}$ measure, *Top-M* F_1 measure and *Precision* measure for a given annotation length M. In our case, we set M to 5. F_β measure is defined as $(1+\beta^2)\cdot(P\cdot R/\beta^2 P+R)$, where P is the averaged per-image precision and R is the averaged per-image recall. When we set β to 0.5, we put more emphasis on precision than recall. The reason is that the ground-truth annotation length is usually more than the fixed length we used for most of the images. Therefore, even we get all the annotations correct, we still cannot reach the best recall score. Instead, we look for better performance by considering the true positives in the total five annotations. However, to give more information on the performance, we also provide the results evaluated by standard F_1 measure and $Precision$ measure.

- Image retrieval: The performance is evaluated by the ranks of the relevant images in the returned results. We have five human assessors launched queries using each database image and provide relevance information on the retrieved images. The degree of relevance of a retrieved image is calculated by the total number of assessors who submit "relevant" decision divided by five. Further statistical evaluation relies on the standard image retrieval measure: *Mean average precision of top D retrieved images* over all the images. Let D be the retrieved image set and R be the relevant ones with size $|R|$. Given a query Q, the average precision is defined as $AP(Q) = \frac{1}{|R|}\sum_{i=1}^{|R|}\frac{i}{Rank(R_i)}$, and the mean average precision (MAP) is the averaged AP over all the images.

3.3.4 Co-occurrence Pattern Detection Results

3.3.4.1 Experiment I: Co-occurrence measure study

We apply our co-occurrence pattern detection approach on a network built from the training set of each dataset. LabelMe contains 2,500 individual concepts, SUN'09 contains 5,800 concepts, and OSR has 520 concepts.

We demonstrate that our combined co-occurrence measure of NGD, NTD, and ALA is more effective than each of the individual measures in co-occurrence network construction as well as co-occurrence pattern detection in the following experiments. First, we compare example pairwise concept co-occurrence scores computed by different measures in Table 3.3. The scores are averaged over the three datasets and normalized to the range $[0, 1]$. Generally, we find the results from NGD, NTD, ALA are more coherent on the pairs with degrees of co-occurrences that are more consistent to human perception (e.g., "mountain-tree", "sky-cloud", and "road-car") than the less consistent ones (e.g., 'sand-sea'", "person-terrance" and "rock-hill"). However, our combined measure is able to reach the maximum consensus among the three. For example, our combined measure is able to leverage the information from NGD and NTD to increase the co-occurrence score of ALA from 0.448 to 0.532 for the pair of "mountain-tree", and is able to use local information from ALA to improve the co-occurrence measure of NGD and NTD for the pair of "wall-staircase". From Table 3.4 we can observe the effectiveness of using the combined measure in co-occurrence pattern detection evaluated by the modularity score (Eq. 3.1). Our combined measure gives the best performance in modularity measure from $5th$ level to $10th$ level in the hierarchy. The reason for this is that the combined measure can leverage both the global and local co-occurrences as well as utilize both the semantic and visual information.

Table 3.3: Pairwise co-occurrence scores for example concept pairs by using NGD, NTD, ALA and the combination of the three.

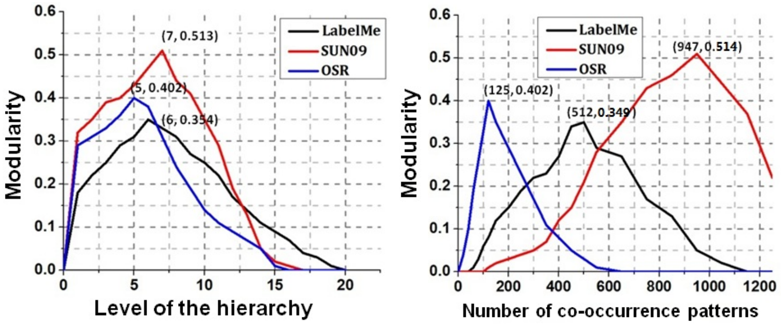
Pairwise Co-occurrence Scores (normalized to [0,1])				
Concept Pairs	NGD	NTD	ALA	Combined
mountain-tree	0.551	0.597	0.448	0.532
sky-cloud	0.713	0.825	0.629	0.722
road-car	0.533	0.614	0.687	0.611
street-building	0.429	0.475	0.512	0.472
sand-sea	0.217	0.483	0.359	0.353
ground-grass	0.261	0.385	0.297	0.314
person-terrance	0.097	0.152	0.219	0.156
door-window	0.483	0.509	0.411	0.468
rock-hill	0.202	0.317	0.384	0.301
sun-land	0.215	0.158	0.278	0.217
river-boat	0.343	0.416	0.357	0.372
sidewalk-sign	0.294	0.187	0.371	0.284
field-fence	0.482	0.359	0.411	0.417
wall-staircase	0.128	0.119	0.274	0.174
curb-streetlight	0.213	0.319	0.307	0.280

Table 3.4: Averaged modularity scores (Q) from $5th$ to $10th$ level.

Modularity Scores				
Datasets	NGD	NTD	ALA	Combined
OSR	0.218	0.259	0.224	**0.275**
SUN09	0.152	0.170	0.143	**0.212**
LabelMe	0.173	0.164	0.139	**0.197**

3.3.4.2 Experiment II: Impact from the hierarchy level

Figure 3.2(a) shows the change in modularity for different levels of hierarchy in the three datasets. We observe that the maximum of modularity for LabelMe occurs at level 6 with $Q \approx 0.354$, the maximum for SUN'09 occurs at level 7 with $Q \approx 0.513$ and the maximum for OSR occurs at level 5 with $Q \approx 0.402$. This indicates that the individual concepts in SUN'09 have significant community property than OSR and LabelMe, and even appear at lower level of SUN'09 (from level 7 to level 12), the community property is comparatively large compared to the LabelMe and OSR datasets. Figure 3.2(b) shows the correspondence between the number of co-occurrence patterns and the modularity values at different levels of the hierarchical community structures. From Figure 3.2(b) we can compute the average number of concepts in the co-occurrence patterns by dividing the total number of concepts by the number of co-occurrence patterns. LabelMe has approximately 5 concepts averaged over all the co-occurrence patterns at the maximum modularity point, similarly, SUN09 has 6 concepts and OSR has 4 concepts. Note the averaged number of concepts in the co-occurrence patterns are consistent with the averaged number of concepts contained in the training images.



(a)

(b)

Figure 3.2: (a) Modularity vs. level of the hierarchy. (b) Modularity vs. the number of co-occurrence patterns.

3.3.5 Automatic Image Annotation Results

3.3.5.1 Experiment I: Co-occurrence measure study

Table 3.5 presents the precisions obtained for the three datasets at different annotation length (*Pre*@1, *Pre*@3, *Pre*@5, *Pre*@10) by using four co-occurrence measures: NGD, NTD, ALA, and our *Combined*. *Pre*@N denotes the precision of annotations in the first N words using 60% percent of the dataset for training. Overall, our combined co-occurrence measure achieves the best performance especially when the annotation length is larger than 1. The reason is that for more annotations more co-occurrence information can be utilized. Generally, when the length of the annotation becomes larger, it deteriorates the annotation precision, however, using combined co-occurrence information our proposed measure still can achieve relatively stable performance regardless of the dataset complexity differences. Furthermore, the number of true positives exceeds 30% for our co-occurrence measure at the length of ten annotations which implies that at least three annotations on average are correctly given by our approach. Note that, in general, the contribution from local visual co-occurrence, which is adopted by ALA, surpasses the contributions from global semantic co-occurrence and global visual co-occurrence which are adopted by NGD and NTD, respectively. This demonstrates that each dataset has unique co-occurrence patterns which are different from the global ones. However, by introducing the global information, we can actually consolidate the common patterns which may lack enough samples in a local dataset and weaken the unusual patterns.

3.3.5.2 Experiment II: Annotation performance

To demonstrate the effectiveness of our proposed framework for the image annotation application, we evaluate the following approaches as shown in Table 3.6:

- **Baseline-Gen model**: Our generative implementation for individual concept inference unified with concept signature representation served as the base model. (The base model does not include co-occurrence pattern detection and random walk boosting).

- **Baseline-Dis model** The discriminative version of the baseline-gen model. The other setup is the same as in baseline-gen.

- **CRF**: The conditional random field (CRF) based image annotation approach by Xiang et al. [178] that uses the original pairwise co-occurrences from a network structure without hierarchical co-occurrence pattern detection. We re-implemented it to compare it with our hierarchical pattern scheme.

- **Context**: The object detection and localization approach by Choi et al. [27] that is used for image annotation. They introduced a tree-structured context model which is comparable to our network structure and hierarchical patterns. We re-implemented it to compare its performance with our approach.

Table 3.5: Precisions at different annotation lengths by using different co-occurrence measures.

LabelMe				
Co-occurrence Measure	Pre@1	Pre@3	Pre@5	Pre@10
NGD	0.3393	0.2584	0.2230	0.1276
NTD	0.3752	0.2772	0.2481	0.2025
ALA	0.3806	0.2857	0.2564	0.1847
Combined	**0.4628**	**0.4533**	**0.4279**	**0.3104**
SUN09				
Co-occurrence Measure	Pre@1	Pre@3	Pre@5	Pre@10
NGD	0.3528	0.2693	0.2432	0.1384
NTD	0.3423	0.2537	0.2593	0.1457
ALA	0.3516	0.2714	0.2581	0.1543
Combined	**0.4332**	**0.4233**	**0.4017**	**0.3042**
OSR				
Co-occurrence Measure	Pre@1	Pre@3	Pre@5	Pre@10
NGD	0.3393	0.2584	0.2230	0.1276
NTD	0.3752	0.2772	0.2481	0.2025
ALA	0.3806	0.2857	0.2564	0.1847
Combined	**0.4423**	**0.4323**	**0.4264**	**0.3504**

- **HCP-Gen**: This is our proposed framework integrating generative concept inference, co-occurrence pattern and random walk boosting. HCP refers to hierarchical co-occurrence pattern.
- **HCP-Dis**: A framework with a discriminative concept inference model and everything else is the same as in HCP-Gen.

We evaluate the impact of the training set size by Top-5 $F_{0.5}$ measure averaged over all the testing images. We split the datasets into training and testing sets with three size configurations. For each split configuration we repeated the experiment 10 times by using each of the approaches. Table 3.6 summarizes the data splits, mean performance and standard deviations. The tables show that the impact of training set size is obvious and consistent across different datasets. The larger the training set, the better performance can be achieved for all the approaches. Our approach shows clear improvements over the other models reflected by the maximum % gain (achieved by using HCP-Gen or HCP-Dis). Also, there is a significant performance gain when the training data size exceeds the testing data size for all the three datasets (see the last two columns for each dataset in Table 3.6. In general, all the approaches require at least 50% of the dataset used for training to have reasonable annotation performance. Even the performance of our framework is deteriorated when the training data is under 40%.

Next, to analyze the scalability of our approach, we compare the results on the three datasets with increased complexity (OSR $<$ SUN09 $<$ LabelMe) evaluated by the total number of concepts in the datasets and the number of concepts per image. Table 3.6 shows that generally when the images are complex the performance of the approaches drop. This is demonstrated by the Top-5 $F_{0.5}$ measure. In particular, we observe that our approach achieves better maximum performance gain when the images

Table 3.6: The Top-5 $F_{0.5}$ score and the standard deviation (show in the parenthesis) of automated annotation with different training set sizes.

	LabelMe Dataset [140] (%)			SUN09 Dataset [27] (%)			OSR Dataset [120] (%)		
Methods / % of training data	40%	60%	80%	40%	60%	80%	40%	60%	80%
Baseline-Gen	21.85 (3.253)	29.57 (2.857)	32.81 (2.772)	23.44 (3.157)	32.52 (2.684)	35.14 (2.435)	25.81 (2.217)	35.33 (1.936)	40.47 (1.854)
Baseline-Dis	21.51 (2.857)	31.27 (2.864)	33.03 (2.513)	21.73 (2.679)	33.03 (2.324)	35.94 (2.185)	23.33 (1.906)	34.52 (1.873)	39.72 (1.535)
CRF [178]	25.59 (2.095)	33.81 (2.137)	36.04 (2.241)	26.93 (1.958)	35.71 (1.742)	40.15 (1.699)	27.93 (1.732)	38.91 (1.589)	43.93 (1.489)
Context [27]	26.33 (1.964)	34.13 (1.842)	36.23 (1.765)	27.71 (1.753)	36.58 (1.689)	40.81 (1.626)	28.14 (1.541)	38.63 (1.439)	44.18 (1.387)
HCP-Gen (This paper)	28.74 (1.154)	39.92 (1.112)	41.37 (1.037)	29.52 (1.096)	39.11 (0.965)	44.32 (0.854)	28.71 (0.896)	42.64 (0.859)	**48.36** (0.791)
HCP-Dis (This paper)	28.13 (1.032)	40.85 (1.006)	**43.46** (0.987)	28.23 (1.043)	40.71 (0.958)	**45.68** (0.875)	28.47 (0.955)	41.92 (0.890)	47.71 (0.873)
maximum % gain over CRF	12.30%	**20.82**%	**20.59**%	9.61%	**14.00**%	**13.77**%	2.79%	**9.59**%	**10.08**%
maximum % gain over Context	9.15%	**19.69**%	**19.96**%	6.53%	**11.29**%	**11.93**%	2.03%	**10.38**%	**9.46**%

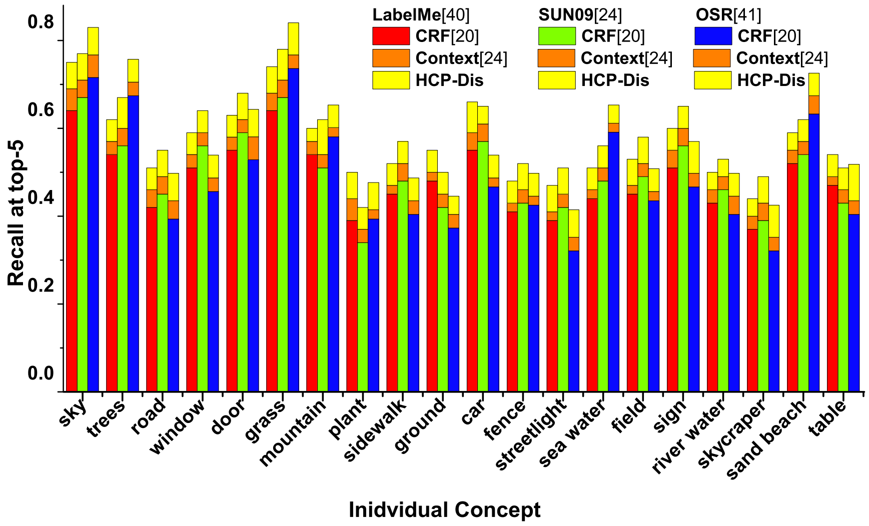


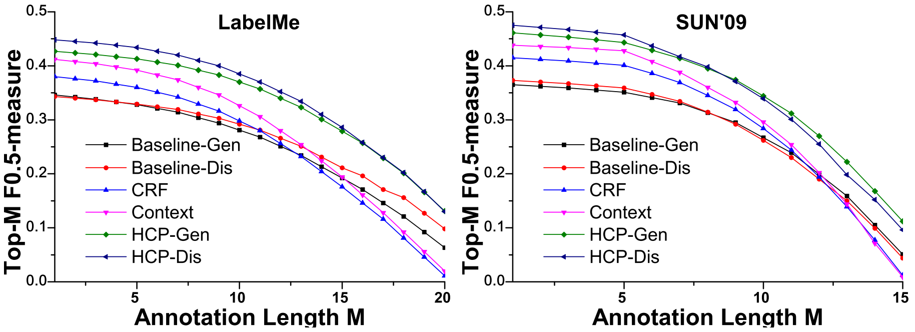
Figure 3.3: The recall rate of common concepts in the three datasets.

have higher complexities. For example, LabelMe usually has more than 10 concepts in an image, the maximum performance gain reaches 20.59% when the training set contains 80% of the images. SUN09 contains on average 5-10 concepts per image, the maximum performance gain is between 11.29% and 14.00%. OSR has the least number of concepts in an image, and the maximum gain is the lowest as well which is approximately 10.00% only. This indicates that our approach is well suited for understanding images with complex scenes. Table 3.6 also shows that the performance increase by our approach is less compared to other approaches when the images are relatively simple as in the OSR dataset.

We further compare the recall rates at top-5 annotation length obtained by CRF [178], Context [27] and our HCP-Dis approach on selected common concepts across the three datasets. The results are given in Figure 3.3. We observe that the contextual information from the three datasets have different effects on individual concept inference. For example, the recall rates for most of the concepts in LabelMe are relatively lower than

SUN09 and OSR. The reason for this is that there are more noisy annotations, such as the misspelling and meaningless words in LabelMe from the folksonomy-style annotations, and these noisy annotations deteriorate the co-occurrence pattern detection performance and have adverse impact on the individual concept refinement. OSR dataset has larger recall rates on outdoor concepts while has smaller recall rates on other concepts. We stack the recall rates obtained by different approaches into a single column and we observe that Context [27] (with hierarchy) performs better than CRF [178] (without hierarchy) while our approach always has the highest performance gain on the recall rate. This demonstrates the effect of using hierarchical co-occurrence patterns vs. no hierarchy. Additionally, the recall rates of CRF [178], Context [27] highly depend on the visual consistency of the semantic concepts. For concepts have large intra-concept visual variations (e.g., "road", "ground", "streetlight", and "skycraper" in Figure 3.3), the performance drops greatly especially for CRF which only considers the original pairwise concept co-occurrences. On the other hand, our approach can maintain relatively stable performance which demonstrates the effectiveness of utilizing contextual information obtained from the detected co-occurrence patterns.

Figures 3.4(a), (b) show the performance comparison based on the Top-M $F_{0.5}$-measure for the three datasets as a function of the annotation length M. As the number of annotations increases, we observe that the performance of baseline approaches, CRF [178] and Context [27] drops faster than our proposed HCP approaches, which demonstrates that our co-occurrence pattern and refinement has a boosting effect on individual concept inference. Further, our approach is more effective in using contextual information than CRF [178] and Context [27] because we explore the correlations of concepts beyond pairwise relationships. We also observe that our discriminative model and generative provide approximately the same boost in performance compared to the other



(a)

(b)

Figure 3.4: (a),(b) show the image annotation performance of the approaches applied to the three datasets measured by Top-5 $F_{0.5}$-measure with annotation length $M = 5$.

approaches. However, HCP-Dis performs better than HCP-Gen for the datasets such as LabelMe and SUN'09 that have more complex scenes and more semantic concepts in a single image. Therefore, we conclude that HCP-Dis has a stronger discriminative power when the number of semantic concepts that share increasingly high visual similarity in an image. Also HCP-Gen can better tolerate the intra-concept visual variation in simple scenes.

Figure 3.6 shows the top-5 annotation results for some example images that are produced by our approach. The annotations in green color are the correctly predicted labels and red ones are mistakenly predicted. It is interesting to look at the annotations in blue. These concepts are inferred from the detected individual concepts and co-occurrence patterns. Although they are not exactly the same as the annotations in the ground-truth, but they are close in the meaning for a specific scenario, e.g., "road" and "path" in an "outdoor - street view" scenario, "people" and "pedestrian" in an "indoor - hall" scenario. This shows that our proposed approach can effectively enrich the annotations by considering the scene concepts implicitly contained in the co-occurrence patterns. The refinement capacity of our approach can be seen from the annotation results of the right image in the second row and left image in the last row where the

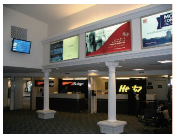
LabelMe Image	Our approach	Ground-truth	LabelMe Image	Our approach	Ground-truth
	Building Sign Trees Sky Road	Carside Clock Tower Building Sign Sky Bicycle Trees Plants Person Walking Path Wall		Floor Window Light Wall People	Pedestrian Door Ceiling Floor Window Wall Plant Sign Corridor Light Trash can Doorway
SUN'09 Image	**Our approach**	**Ground-truth**	**SUN'09 Image**	**Our approach**	**Ground-truth**
	Sky Sign Road Car Trees	Sky Highway Text Car Fence Mountain Trees Sign Car Occluded		Column Floor Wall Table Sign	Screen Column Ceiling Chair Text Check-in-desk Person Occluded Suitcase Wall Floor
OSR Image	**Our approach**	**Ground-truth**	**OSR Image**	**Our approach**	**Ground-truth**
	Trees Building Road Car Sidewalk	Building Cannon Trees Pedestal Sidewalk Staris Road Plant Bus Window Garden Path Person Standing		Trees Sky Cloud Ocean Sand	Sky Trees Mountain Stone Sea water Rock Ship

Figure 3.5: The annotations for the test images from the three datasets by our approach. They are compared with the ground-truth. Green labels are correctly predicted, red ones are wrongly predicted and blue ones have very close semantic meaning to the ground-truth.

ground-truth concepts "check-in-desk" and "bus" are occluded in the image and the similar concept "table" and "car" are enriched by our proposed refinement strategy.

3.3.6 Concept-based Image Retrieval Results

3.3.6.1 Experiment I: Co-occurrence measure study

Table 3.7 gives the mean average precisions (MAP) for the datasets at four different sizes of retrieved images (MAP@5, MAP@10, MAP@15, MAP@20) by using four co-occurrence measures: NGD, NTD, ALA, and combined. MAP@N represents the mean average precision of retrieved images in the size of N using 60% percent of the dataset for training. The results in Table 3.7 show that our combined co-occurrence measure achieves the best performance at all sizes of the retrieved images.

From Table 3.7 we can observe that the combined co-occurrence measure achieves the best performance and the performance is stable when the size of the retrieved images is less than 15. Even when the size is 20, the combined co-occurrence

Table 3.7: Mean average precision for different sizes of retrieved images by using different co-occurrence measures.

LabelMe				
Co-occurrence Measure	MAP@5	MAP@10	MAP@15	MAP@20
NGD	0.2564	0.2317	0.1869	0.1003
NTD	0.2616	0.2484	0.2195	0.1574
ALA	0.2543	0.2336	0.1752	0.1249
Combined	**0.2825**	**0.2617**	**0.2797**	**0.1809**
SUN09				
Co-occurrence Measure	MAP@5	MAP@10	MAP@15	MAP@20
NGD	0.2646	0.2334	0.1954	0.1172
NTD	0.2476	0.2318	0.2094	0.1290
ALA	0.2584	0.2027	0.1853	0.1274
Combined	**0.2923**	**0.2898**	**0.2517**	**0.1972**
OSR				
Co-occurrence Measure	MAP@5	MAP@10	MAP@15	MAP@20
NGD	0.2738	0.2418	0.1989	0.1373
NTD	0.2864	0.2529	0.2046	0.1508
ALA	0.2953	0.2591	0.2153	0.1643
Combined	**0.3394**	**0.3004**	**0.2846**	**0.2038**

measure can still have reasonable results in all three datasets. Note that, in general, the contributions from the three individual measures are relatively the same for all sizes of retrieved images. But the boost in MAP values is clear when combining the three measures. This demonstrates that the co-occurrence information from the three measures will compensate each other and it is helpful in learning more accurate concept relationships. Note that the MAP measure is affected by two factors: the difficulty of the dataset and the number of retrieved images. The combined measure can achieve a better MAP compared to the individual measures for all datasets of varying difficulty levels and retrieved image sizes.

3.3.6.2 Experiment II: Image retrieval performance

The goal is to show the effectiveness of our concept inference framework for image retrieval task. We implement and evaluate the following approaches for comparison as summarized in Table 3.8. We also vary the training set size to show its impact on the retrieval performance.

- **Baseline-I**:The content-based image retrieval framework that compares the image similarity by directly computing the Euclidean distance between the visual feature vectors as described in Section 4.2.

- **Baseline-II**: The proposed framework integrated with SVM-based individual concept inference. The concept signatures are used directly without refinement by co-occurrence patterns.

- **Semi-Supervised graphical model (SSG)**: The approach in [186] uses a latent-tree to find the relationship between semantic concepts. The pairwise relevance is obtained from the graphical model directly. No hierarchical co-occurrence patterns

Table 3.8: Mean average precision of top-10 retrieved images with different training set size.

	LabelMe Dataset [140] (%)			SUN09 Dataset [27] (%)			OSR Dataset [120] (%)		
Methods / % of training	40%	60%	80%	40%	60%	80%	40%	60%	80%
Baseline-I	7.64 (2.857)	11.93 (3.383)	14.82 (2.754)	9.46 (2.953)	13.53 (2.714)	15.74 (2.906)	15.82 (2.186)	25.37 (2.346)	28.27 (1.974)
Baseline-II	14.43 (2.952)	21.58 (2.742)	26.03 (2.563)	12.71 (3.126)	18.14 (3.064)	22.17 (2.547)	18.93 (2.836)	25.12 (2.914)	29.79 (2.464)
SSG [186]	17.78 (2.532)	24.61 (2.734)	26.94 (2.513)	19.53 (2.631)	25.71 (2.345)	28.45 (2.194)	21.22 (3.126)	28.54 (2.432)	31.82 (1.987)
HSI [33]	17.93 (2.964)	25.17 (2.347)	27.27 (2.146)	19.71 (3.156)	26.15 (2.343)	28.61 (3.134)	21.78 (2.432)	28.78 (1.524)	32.15 (1.768)
HCP-IR (This paper)	18.96 (1.532)	28.97 (1.123)	**32.47** (1.233)	21.06 (1.518)	30.17 (1.425)	**34.22** (1.236)	23.11 (1.435)	33.46 (1.346)	**37.99** (0.983)
% gain over SSG	6.64%	**17.72**%	**20.53**%	7.83%	**19.86**%	**20.28**%	8.91%	**17.23**%	**19.39**%
% gain over HSI	5.74%	**15.10**%	**19.07**%	6.85%	**15.37**%	**19.60**%	6.11%	**16.26**%	**18.16**%

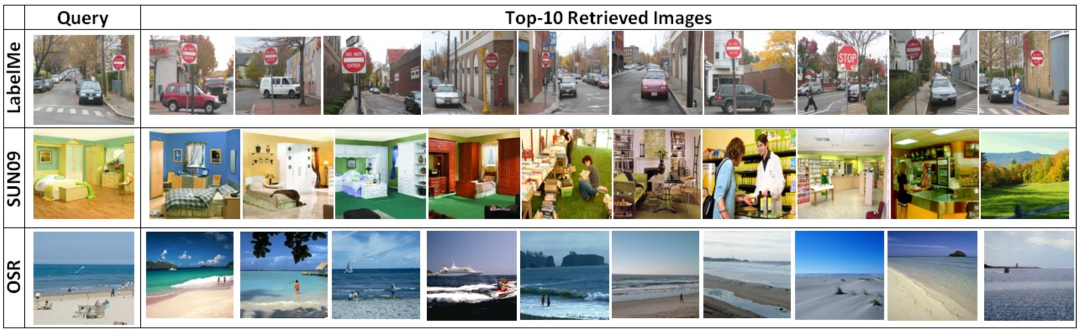


Figure 3.6: An example of the top-10 retrieved images by our proposed approach. The retrieved images are ranked based on their semantic distance to the query. The top row shows the correctly retrieved images with street view and the stop sign. In the middle row, the top retrieved images correctly match the bedroom scene represented in the query. And in the last row, the images with a beach scene and people are placed at the top positions.

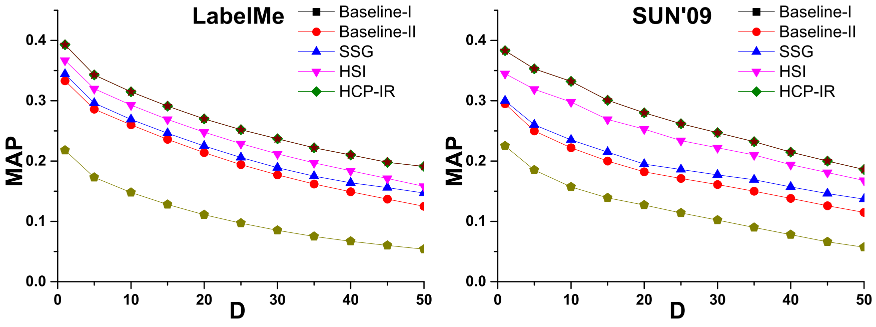
are detected.

- **Hierarchical semantic indexing (HSI)**: The retrieval framework proposed in [33] uses the information from generated hierarchical semantic relationships between concepts for comparing image similarity. However, as compared to our work, they do not consider the co-occurrence between concepts, and their concept distance is defined on WordNet.

- **HCP-IR**: Our proposed approach integrated with hierarchical co-occurrence pattern detection and concept signature refinement. We implemented the discriminative model here.

We repeat the split of each of the dataset for ten times. From Table 3.8 we can observe that the larger the training set size for all the three datasets, the larger MAP can be achieved by all the approaches. The standard deviations are also given in this table. Baseline-I achieves the worst performance which concludes that traditional content-based image retrieval paradigm is not suitable for retrieving images containing many semantic concepts with a large visual variations. SSG is only marginally better than our Baseline-II approach, for the reason that it only considers the pairwise relationship between individual concepts and the approach is not intended to use images from complex scenes. HSI outperforms SSG while our HCP-IR significantly outperforms both SSG and HSI by 5.74%-20.53%. This result validates our assumption that the proposed hierarchical concept co-occurrence patterns can boost the individual concept inference. In particular, we can observe that when using only 40% of the dataset for training, our method can still achieve comparatively good performance than SSG and HSI. An example of the retrieval results by using our HCP-IR approach with 80% training data for the three datasets is shown in Figure 3.6. We can observe that the returned images are more semantically related to the scene concept reflected in the query images rather than just visually related. The overall performance of all the approaches decrease when the dataset becomes more complex. However, our approach can maintain a stable maximum gain over SSG [186] and HSI [33].

Figures 3.7(a), (b) summarize the results for MAP at top-D retrieval results. Our model (HCP-IR) consistently outperforms the other approaches with varying number of retrieved images on the three datasets. This shows the effects of semantic concept correlations and the concept signature descriptor in the context of image retrieval. The results demonstrate that all the components of our framework are essential: (1) detecting individual semantic concepts is important for retrieving images of complex scenes (LabelMe, SUN'09) as Baseline-II is more effective than Baseline-I (directly using low-level features without semantic learning). (2) learning more sophisticated concept correlation models (HSI, HCP-IR) improves performance over simple pairwise relationships (SSG). We also note a higher precision for OSR than for the other two datasets. This is due to a relatively small number of individual concepts present in the dataset, and therefore, the detected co-occurrence patterns are more significant in more compact forms.



(a)

(b)

Figure 3.7: (a), (b) show the image retrieval performance of the approaches applied to the three datasets measured by Top-D MAP with varied number of retrieved images D.

3.4 Conclusions

This paper has made a novel contribution to the literature on context-based co-occurrences in computer vision where co-occurrences of concepts are used as contextual cues for improved concept inference. It introduced a framework for individual concept inference and refinement by exploring the concept co-occurrence patterns in images with network community detection algorithms. The framework is evaluated for automated image annotation and concept-based image retrieval tasks using the new concept signature representation. The approach is tested on recent practical datasets and compared with the state-of-the-art methods. The experimental results convincingly show the following: (a) The importance of the hierarchy of co-occurrence patterns and its representation as a network structure, (b) The effectiveness of the approach for building individual concept inference models and the utilization of co-occurrence patterns for refinement of concept signature as a way to encode both visual and semantic information. In the future we will explore the message-passing approach for concept signature refinement and compare it with the random walk based approach.

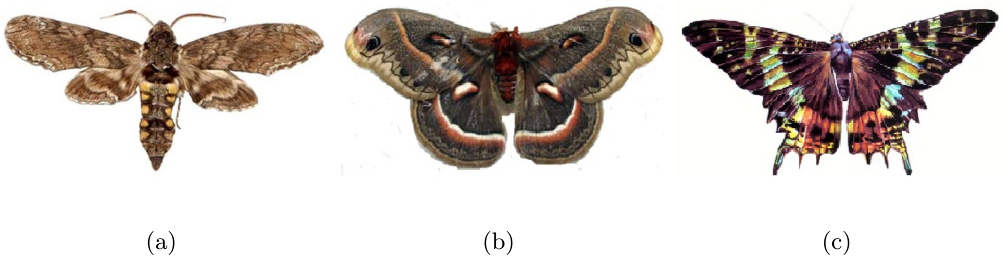
Chapter 4

Automated Moth Species Identification and Retrieval

4.1 Introduction

Moths are important life forms on the planet with approximately 160 000 species discovered [21], compared to 17 500 species of butterflies [21], which share the same insect Order with Lepidoptera. Although most commonly seen moth species have dull wings (e.g., the Tomato Hornworm moth, see Figure 4.1(a)), there are a great number of species that are known for their spectacular color and texture patterns on the wings (e.g., the Giant Silkworm moth and the Sunset moth, see Fig. 1b and Fig.1c respectively). As a consequence, much research on identifying the moth species from the entomologist side has focused on manually analyzing the taxonomic attributes on the wings such as color patterns, texture sizes, spot shapes, etc., in contrast with the counterpart biological research that classifies species based on DNA differences.

As image acquisition technology advances and the cost of storage devices decreases, the number of specimen images in entomology is grown at an extremely rapid



(a) (b) (c)

Figure 4.1: Moth wings have color and texture patterns at different levels of complexity based on their species: (a) Tomato Hornworm, (b) Giant Silkworm and (c) Sunset. Photo courtesy of Google Image search engine.

rate both in private database collections and over the web [18, 19, 87]. Species identification, relying on manually processing images by entomologists and highly trained experts, is time-consuming and error-prone. The demand for more automated and efficient methods, to meet the requirements of real world species identification such as agriculture and border control, is increasing. Given the lack of manually annotated text descriptors to the images, and the lack of consensus on the annotations caused by the subjectivity errors of the human experts, engines for archiving, searching and retrieving insect images in the databases based on keywords and textual metadata face great challenges in feasibility.

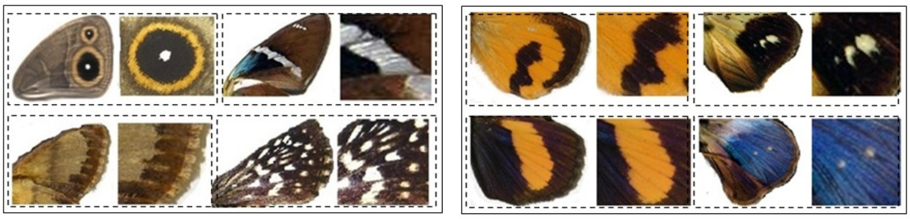
The progress in computer vision and pattern recognition algorithms provides an effective alternative for identifying the insect species and many computer assisted systems that incorporate these algorithms have been invented in the past two decades [39, 60, 81, 144, 176]. In the image retrieval domain, one of the common approaches introduced to complement the difficulties in text-based retrieval relies on the use of Content-Based Image Retrieval (CBIR) systems [20, 149, 190], where sample images are used as queries and compared with the database images based on visual content similarities [11, 169] (color, texture, object shape, etc.). In both the identification and retrieval scenarios, visual features that are extracted to represent morphological and taxonomic information play an important role in the final performance. Context information is often used to help improve individual detection performance of the visual features [38].

These intelligent systems provide a number of attractive functions to entomologists, however, drawbacks have been revealed in several aspects:

- First, most systems only extract visual features that do not contain any *semantic* information. However, recent research [73] shows that human users are more expecting to access images at the *semantic* level. For example, users of a system are more likely to *find all the moths containing eye spots on the dorsal hind wings* rather than to *find all the moth containing a dark blue region near the bottom of the image.* An intermediate layer of image semantic descriptor that can bridge the gap between user information need and low-level visual feature is absent in most existing systems.

- Second, most systems involve no human interaction and feedback. For example, the insect classification system introduced by L. Zhu et al. [199] works in an autonomous way on feature selection and classification. The retrieval systems [11, 168, 169] for butterfly images do no ask users to provide feedback and refine the results on the fly. However, the need for user-in-the-loop stems from the fact that intelligent systems are not smart enough to interpret image in the same way as humans. For example, two different species could be identified as the same based on their visual similarity. Without human intervention, the system will not be able to tune its parameters and correct the mistakes.

- Third, the current systems for species identification overlook the co-occurrence relationship among features. For example, in [85, 86, 106, 176], the co-occurrence of features as contextual cues was not investigated to reduce or even remove the uncertainty in species identification. Intuitively, such information is helpful to better distinguish insect species. For example, in some species of Lepidoptera, a border “eye spot” feature



(a) (b)

Figure 4.2: Sample moth wings illustrate the Semantically-Related Visual (SRV) attributes. (a) Four sets of SRV attributes on the dorsal fore wings: eye spot (top left), central white band (top right), marginal cuticle (bottom left) and snowflake mosaic (bottom right). In each set, the right image is the enlarged version of the left image. (b) Four sets of SRV attributes on the ventral hind wings. Note it is harder to described the images in a semantic way with simple texts compared to the images in group (a).

may often be accompanied with a central "bands" feature on the wings, while other species may not have this combination of wing features. Such co-occurrence of features could be very useful to improve the performance of species identification.

4.1.1 Contributions of This Chapter

In this chapter, we present a new system for automated moth identification and retrieval based on the detection of visual attributes on the wings. The objective of our method is to mimic human behavior on differentiating species by looking at specific *visual contexts* on the wings. More specifically, the notion of "context" refers to discovering certain attribute relationships by taking into account their co-occurrence frequencies. The main motivation of our system relies on the conjecture that the attribute co-occurrence patterns encoded on different species can provide more information for refining the image descriptors. Unlike earlier works, we attempt to address all the above mentioned problems, and the contributions of this paper are summarized as follows:

1. We build image descriptors based on so-called *Semantically Related Visual (SRV) attributes*, which are the striking and stable physical traits on moth wings. Compared

to traditional visual features used in many systems, our SRV attributes have human-designated names (e.g., blue preapical spot, white central bands, yellow eye spot, etc.) which makes them valuable as semantic cues. Some examples of SRV attributes are shown in Fig 4.2. The probabilistic existence of these attributes can be discovered from images by trained detectors using computer vision and pattern recognition techniques. Compared to traditional image feature representations, which is usually a vector of numeric values denoting the significance of visual properties, such as the curvature of a shape boundary, the color intensity of a region, etc., the SRV attribute based image descriptor provides a semantically rich way which is much closer to the way that humans describe and understand images.

2. Our system detects and learns SRV attributes in a supervised way. The SRV-attributes are manually labeled by human experts to a small subset of the image database that is used for training the attribute detectors. The core of the detector is a probabilistic model that can infer SRV-attribute occurring scores from the unlabeled testing images. We characterize individual images by stacking the probabilistic scores of the present SRV attributes into a so-called *SRV-attribute signature*. The species identification and retrieval tasks are performed by comparing the SRV-attribute signature similarity. Specifically, in the image retrieval task, we incorporate human relevance feedback scheme (often collected via user click-and-mark data) with the goal of retrieving more relevant images in future search sessions. We also consider ranking results based on constraints of multi-attribute queries and the relative strengths to improve the effectiveness of attribute based image search.

3. We explicitly explore the co-occurrence relationship of SRV attributes. The underlying idea is that the attributes that appear together frequently across many images

are likely to form a certain pattern. Moths from the same species often exhibit consistent patterns of SRV attributes on the wings. In this paper, we propose a novel approach that utilizes the external knowledge from human labeling in the training set to build a co-occurrence network of SRV attributes and further uncover the patterns of these attributes and use them as contextual cues to improve the individual attribute detection performance.

Our experimental evaluation shows that the proposed SRV attribute based image representation can improve moth species identification accuracy and image retrieval precision considering different datasets. Experimental results also demonstrate that the proposed system can outperform state-of-the-art systems in the literature [150, 169] in terms of effectiveness. We also evaluate other aspects of the proposed system (such as the impact of parameters) in the experiment section.

4.2 Technical Details

4.2.1 Moth Image Dataset

The dataset used in this study is collected from an online library of moth, butterfly and caterpillar specimen images created by Dr. Dan Janzen [80] over a long-term and ongoing project started in 1977 in northwestern Costa Rica. The goal of the inventory is to have records for all the $12,500+$ species in the area. As of the end of 2009, the project had collected images of $4,500$ species of moths, butterflies and caterpillars. We use a subset of the adult moth images under the permission of Dr. Dan Janzen. The dataset is publicly available at http://janzen.sas.upenn.edu.

The images are available for both the dorsal and ventral aspects of the moths. Each image was resized into 600×400 pixels in resolution, and is in RGB colors. Our

Table 4.1: Families, species and the number of samples in each species used in our work.

Sub-Families	Species	Images	Sub-Families	Species	Images
Catolacinae	*Ceroctenaamynta*	101	*Nystaleinae*	*Bardaximaperses*	74
Catolacinae	*Eudocimamaterna*	85	*Nystaleinae*	*Dasylophiabasitincta*	78
Catolacinae	*Eulepidotisfolium*	76	*Nystaleinae*	*Dasylophiamaxtla*	98
Catolacinae	*Eulepidotisrectimargo*	57	*Nystaleinae*	*Nystaleacollaris*	85
Catolacinae	*Hemicephalisagenoria*	121	*Nystaleinae*	*Tachudadiscreta*	112
Catolacinae	*Thysaniazenobia*	79	*Pyrginae*	*Atarnessallei*	101
Dioptinae	*Chrysoglossanorburyi*	75	*Pyrginae*	*Dyscophellusphraxanor*	86
Dioptinae	*Erbessaalbilinea*	98	*Pyrginae*	*Tithraustesnoctiluces*	96
Dioptinae	*Erbessasalvini*	117	*Pyrginae*	*Entheusmatho*	99
Dioptinae	*Nebulosaerymas*	69	*Pyrginae*	*Hyalothyrusneleus*	82
Dioptinae	*Tithrausteslambertae*	87	*Pyrginae*	*NascusBurns*	94
Dioptinae	*Polypoetesharuspex*	92	*Pyrginae*	*Phocidesnigrescens*	104
Dioptinae	*Dioptislongipennis*	92	*Pyrginae*	*Quadruscontubernalis*	69
Hesperiinae	*Methionopsisina*	122	*Pyrginae*	*Urbanusbelli*	88
Hesperiinae	*Neoxeniadesluda*	107	*Pyrginae*	*MelanopygeBurns*	76
Hesperiinae	*SalianaBurns*	70	*Pyrginae*	*Myscelusbelti*	103
Hesperiinae	*Salianafusta*	97	*Pyrginae*	*Mysoriaambigua*	93
Hesperiinae	*TalidesBurns*	70	*Rifargiriinae*	*Dicentriarustica*	78
Hesperiinae	*Vettiusconka*	96	*Rifargiriinae*	*Farigiasagana*	84
Hesperiinae	*Aromaaroma*	135	*Rifargiriinae*	*Hapigiodessigifredomarini*	93
Hesperiinae	*Carystoidesescalantei*	88	*Rifargiriinae*	*Malocampamatralis*	100
Nystaleinae	*Lirimirisguatemalensis*	95	*Rifargiriinae*	*MeragisaJanzen*	65
Nystaleinae	*Isostylazetila*	99	*Rifargiriinae*	*Naprepahoula*	74
Nystaleinae	*Oriciadomina*	101	*Rifargiriinae*	*Pseudodryaspistacina*	83
Nystaleinae	*Scoturaleucophleps*	117	*Rifargiriinae*	*Rifargiadissepta*	69

complete dataset contains $37,310$ specimen images covering $1,580$ species of moth, but a majority of the species have less than twenty samples. Because our feature and attribute analysis are based on regions on the wings, and some specimens show typical damage ranging from age-dependent loss of wing scales (color distortion), missing parts of wings (incomplete image), or uninformative orientation differences in the wings or antennae, this makes the number of qualified samples even less, and we carefully selected fifty species across three family groups and six sub-family groups: *Hesperiidae* (*Hesperinae*, *Pyrginae*), *Notodontidae* (*Dioptinae*, *Nystaleinae*) and *Noctuidae* (*Catolacinae*, *Heterocampinae* [=*Rifargiriinae*]) from the original dataset. This new sub-collection has a total of 4,530 specimens of good quality (see Table 4.1 for the distribution of the species used in our work).

We show sample images of twenty representative species out of the fifty species used in our work in Figure 4.3. The moth specimens have been photographed against an approximate uniform (usually white or grey) background, but often with shadow artifacts. The specimens are curated in a uniformed way with the wings horizontal and generally with the hind margin of the forewing roughly perpendicular to the longitudinal axis, which facilities the subsequent image processing and feature extraction steps.

4.2.2 System Architecture

The flowchart of the proposed moth identification and retrieval system is shown in Figure 4.4. The system architecture contains five major parts: 1) information extraction of moth images, 2) SRV attribute detection on moth wings, 3) co-occurrence network construction and co-occurrence pattern detection for the SRV attributes, 4) image signature building and refinement based on SRV attributes and their co-occurrence patterns, and finally 5) applications in moth species identification and retrieval. We

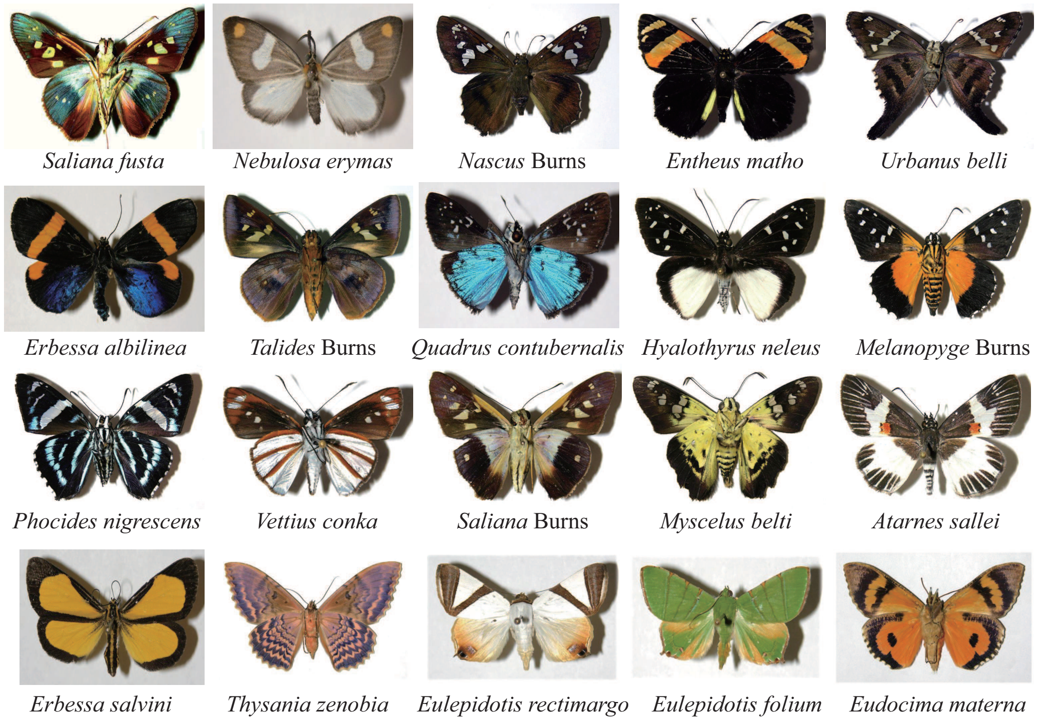
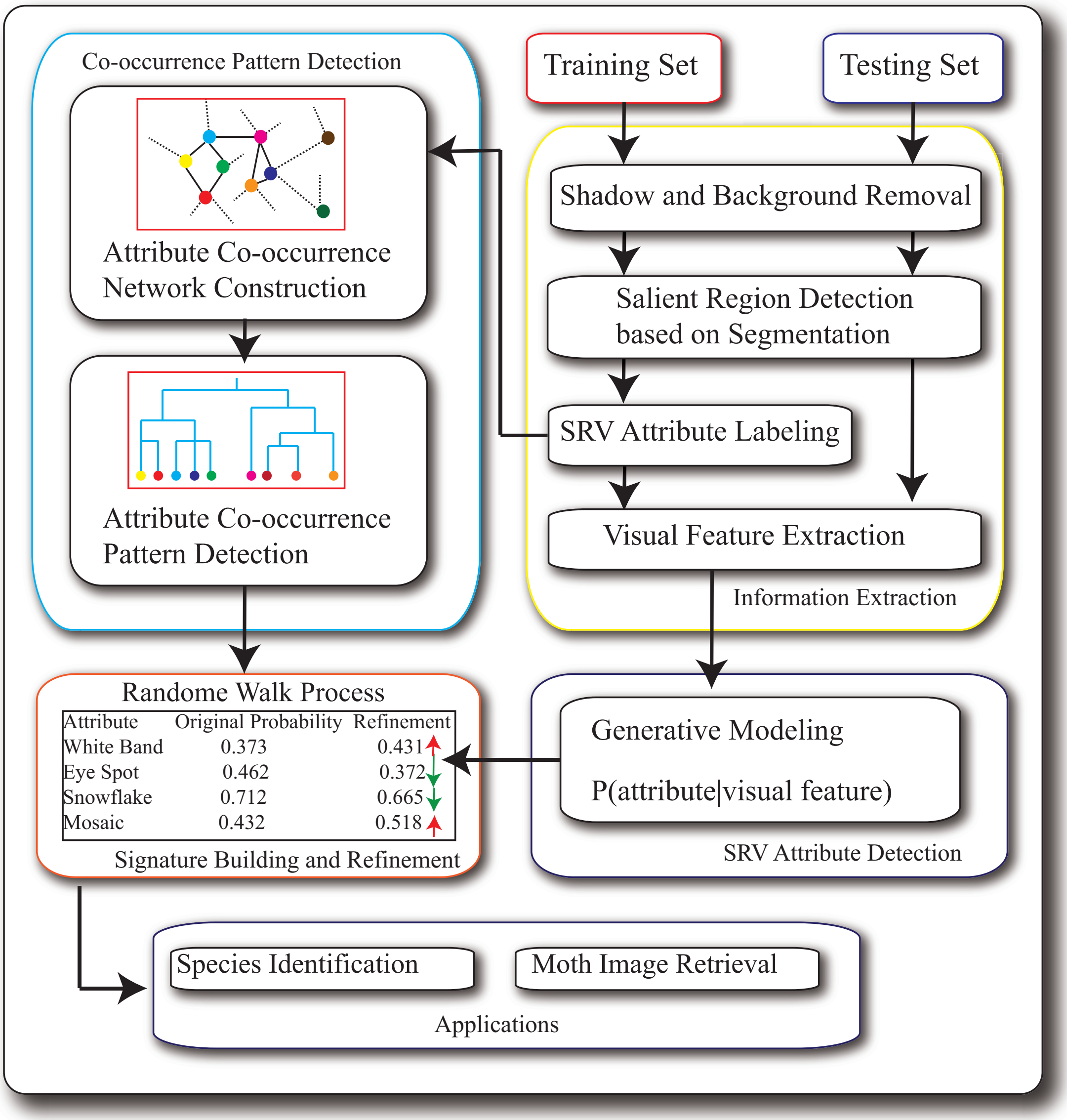


Figure 4.3: Sample images for twenty moth species selected from all the species used in this work. We do not show all the species due to the space limit.

give the details about each part in the following sections.

The information extraction module consists of several steps including background and shadow removal, salient region detection by segmentation, SRV attribute labeling for the training set and visual feature extraction.

In order to train the attribute detectors, we use a small subset of the image collection as the training set. Each training image is segmented manually into regions and the attributes labeled manually to the corresponding regions. The SRV attribute detector is learned from extracted local visual features and the SRV attribute labels by modeling the joint probability of occurrence. After the joint distribution is obtained, we infer the posterior probabilities of attributes from the visual features of the testing images without attribute labeling. The output of the detectors is a pool of the posterior probability scores of each attributes, which is combined into the attribute signature


Co-occurrence Pattern Detection
Training Set
Testing Set
Attribute Co-occurrence Network Construction
Shadow and Background Removal
Salient Region Detection based on Segmentation
SRV Attribute Labeling
Attribute Co-occurrence Pattern Detection
Visual Feature Extraction
Information Extraction
Randome Walk Process
Generative Modeling
P(attribute|visual feature)
Signature Building and Refinement
SRV Attribute Detection
Species Identification
Moth Image Retrieval
Applications

Attribute	Original Probability	Refinement
White Band	0.373	0.431↑
Eye Spot	0.462	0.372↓
Snowflake	0.712	0.665↓
Mosaic	0.432	0.518↑

Figure 4.4: The flowchart of the proposed moth species identification and retrieval system. It consists of: 1) information extraction, 2) SRV attribute detection, 3) attribute co-occurrence pattern detection, 4) attribute signature building and refinement, and 5) moth identification and retrieval applications.

representation of the images.

As the attribute detection relies on the effectiveness of the low-level features to some extent, and in order to improve the detection accuracy by bridging the semantic gap, we propose a novel approach to explore the contextual information of the attributes. Specifically, the co-occurrence pattern recognition module is aimed at uncovering the explicit co-occurrence relationship between attributes in images and utilizing it to further improve the individual attribute detection performance. A random walk process is integrated in this module to maximize the agreement on appearance of individual attributes in an image with respect to co-occurrence.

Relevance feedback is a crucial strategy in image retrieval systems for retrieval result refinement. In our system, we provide the application interface with functions like marking the relevance decisions on the retrieved images. However, as the users of the system may have different levels of professional knowledge, we evaluate their expertise by requiring them to participate in a sample species identification test and authorizing them different levels of permissions to submit feedback based on their scores. The following sections will provide the implementation details of each part shown in Figure 4.4.

4.2.3 Feature Extraction

4.2.3.1 Background removal

It is important to partition the images into "background" and "foreground" because the background usually contains disturbing visual information (such as shadows created by the lighting device, bubbles and dirts on the specimen holder, etc.) that can affect the performance of the detector. We adopted the image symmetry based approach [155] for background and shadow removal. The moth image dataset used in this

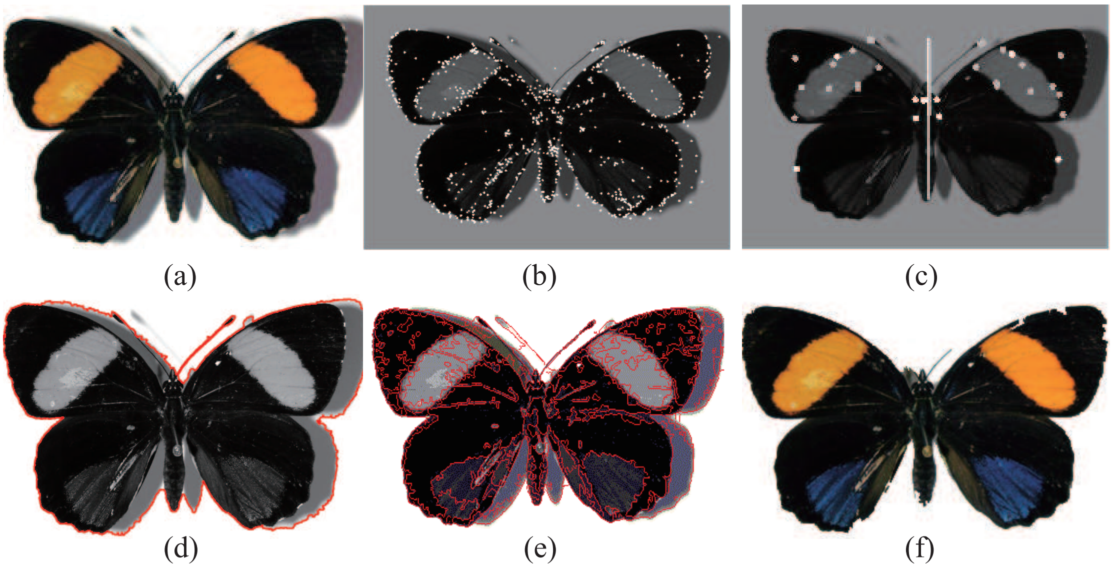


Figure 4.5: Steps for background and shadow removal. (a) Original image (with shadow), (b) Detected SIFT points, (c) Detected symmetry axis, (d) background removed image, (e) segmentation for small parts, and (f) image after shadow removal.

paper have the properties of moth wings with high reflection symmetry (Figure 4.5(a)). Because the shadows have the most salient influence on the following processing steps, and they are not symmetric in the images, we use symmetry as the key constraint to remove the shadow.

The SIFT points of the image are detected (Figure 4.5(b)) and symmetric pairs of the points are used to vote for a dominant symmetry axis (Figure 4.5(c)). Based on the axis, a symmetry-integrated region growing segmentation scheme is applied to remove the white background from the moth body and shadows (Figure 4.5(d)), and the same segmentation process is run with smaller thresholds to partition the image into shadows and small local parts of the moth body (Figure 4.5(e)). Finally, symmetry is used again to separate the shadows from the moth body by computing a symmetry affinity matrix. Since the shadows are always asymmetric with the axis of reflection, their symmetry affinity will have higher values than the parts of moth body, which is used as the criterion to remove the shadows (Figure 4.5(f)).

4.2.3.2 SRV attribute labeling

A sub-region of the moth wing is considered an SRV attribute if: 1) it repeatedly appears on moth wings across many images, 2) it has salient and unique visual properties and 2) it can be described by a set of textual words that are descriptive for the sub-region.

We scan the moth images and manually pick a group of SRV attributes. Similar ways have been utilized for designing “concepts” or “semantic attributes” in image classification and object recognition tasks. For example, building nameable and discriminative attributes with human-in-the-loop [44, 125]. However, compared to their semantic attributes, our SRV attributes cannot be described with concise semantic terms (e.g., “A region with scattered white dots on the margin of the hind wing on the dorsal side”). Therefore, we propose to index the SRV attributes by numbers, e.g., “attribute_1”, “attribute_2” and so forth. We also explicitly incorporate the positions of the SRV attributes into the attribute index. Each moth has two types of wings: the forewing and the hindwing, and each type of wing has two views: the ventral view and the dorsal view, the SRV attribute index is finally defined in an unified format “attribute_No./wing_type/view”, e.g., “attribute_1/forewing/dorsal”, “attribute_5/hindwing/ventral”, etc. Furthermore, as the moths are symmetrical to the center axis, we only label one side of the moth with the index of SRV attributes.

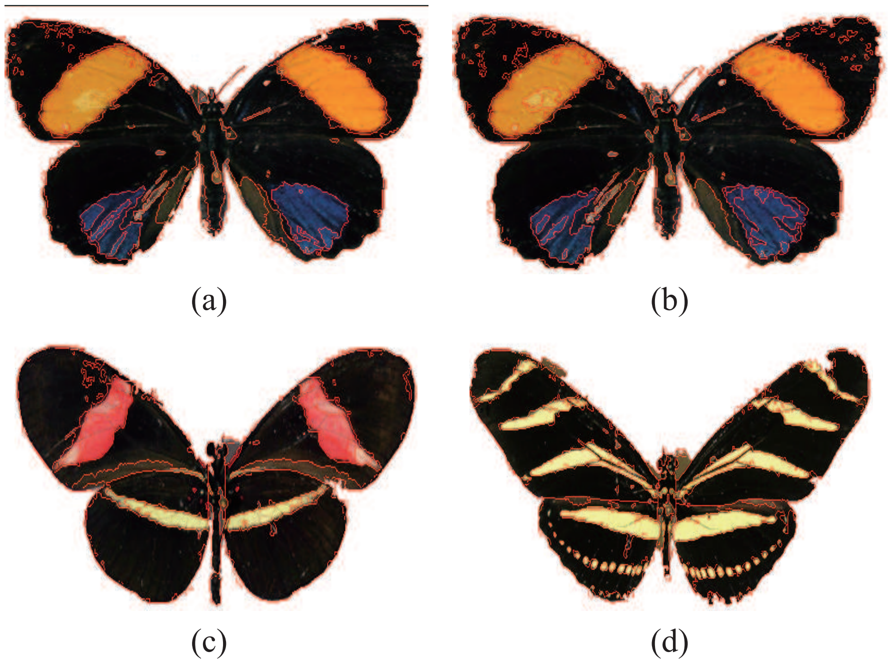
In order to acquire reliable attribute detectors, SRV attributes are labeled by human experts to the regions in the training images. The regions are represented by the minimum bounding rectangles (MBRs) which are produced by using the on-line open source image labeling tool “LabelMe” [141].

4.2.3.3 Salient region detection by segmentation

For the test images, we use the *salient region detector* to extract small regions or patches of various shapes that could potentially contain the interested SRV attributes. A good region detector should produce patches that capture salient discriminative visual patterns in images. In this work, we apply a hierarchical segmentation approach based on reflection symmetry introduced in [155] to jointly segment the images and detect salient regions.

We apply symmetry axis detection on the moth images to compute a symmetry affinity matrix, which represents the correlation between the original image and the symmetrically reflected image. Each pixel has a continuous symmetry affinity value between 0 (perfectly symmetric) and 1 (totally asymmetric), which is computed by the Curvature of Gradient Vector Flow (CGVF) [134]. The symmetry affinity matrix of each image is further used as the symmetry cue to improve the region-growing segmentation. The original region-growing approach considers aggregating pixels into regions by pixel homogeneity. In this paper, we modified the aggregation criterion to integrate the symmetry cue. More details about the approach are explained in [155].

Comparison between Figure 4.7(a) and (b) indicates that by using symmetry, more complete and coherent regions are partitioned. The result in Figure 4.7(b) is obtained by using the same region growing, but without symmetry, so it has many noisy and incomplete regions. The improvements are obtained by using the symmetry cue only. Two more results on salient region detection by using symmetry based segmentation are shown in Figure 4.7(c) and (d).



(a) (b)

(c) (d)

Figure 4.6: Results from salient region detection. (a) symmetry based segmentation, (b) segmentation without using symmetry. Two more results are shown in (c) and (d) by using symmetry based segmentation.

4.2.3.4 Low-level feature extraction

We represent the above detected salient regions by the minimum bounding rectangles (MBRs). The local features of each bounding rectangular are extracted and pooled into numeric vector descriptors. We have three different types of features used to describe each region: a) color-based feature, b) texture-based feature, and c) SIFT keypoint-based feature.

1) HSV color feature. The color feature is insensitive to changes of size and direction of regions. However, it suffers from the influence of illumination variations. For the color feature extraction, the original RGB (Red-Green-Blue) color image is first transformed into HSV (Hue-Saturation-Value) space, and only the hue and saturation components are used to reduce the impact from lighting conditions. We then divide the interval of each component into 36 bins, the image pixels inside the salient region are

counted for each bin, and the histogram of the 72 bins is concatenated and normalized into the final color feature vector.

2) Grey Level Co-occurrence Matrix (GLCM) based texture feature. Texture feature is useful to capture the regular patterns of the spatial arrangement of pixels and the intrinsic visual property of regions. We adopt the gray level co-occurrence matrix (GLCM) proposed by Haralick in [74] to extract the texture features. The GLCM is a pixel-based image processing method.

The co-occurrence matrices in GLCM are calculated based on second order statistics as described in [71]. Each element $P(i, j, d, \varphi)$ in the matrix represents the frequency of co-occurrence of the gray levels of the pixel pair (i, j) along a specific direction φ (e.g., horizontal, diagonal, vertical, etc.) at a distance d (e.g., one to six pixels) between the pixels.

Let $I(x, y)$ denote a two-dimensional digital image of size $M \times N$, and suppose the maximum grey level is G, hence $i, j \in [0, G]$, an element in the GLCM representing the co-occurrence value of two pixels $(x_1, y_1), (x_2, y_2)$ in the image I at angle φ and distance d is expressed in the following equation:

$$P(i, j, d, \varphi) = \sum_{d,\varphi} \Delta[(x_1, y_1), (x_2, y_2)] \tag{4.1}$$

where $\Delta = 1$, if $(x_1, y_1) = i$ and $(x_2, y_2) = j$, else $\Delta = 0$. In the original approach, the author [74] computed 14 statistical features from the matrix. However, the GLCMs can be very sparse, and applying statistics looping through each of the GLCMs can result in a very inefficient procedure given that most of the matrix entries are zero. We use a subset of patches containing the SRV attributes with ground-truth labels. The 14 GLCM features are extracted for each patch. We conduct a classification task for each patch using each of features. The best features that have more discriminative power and lower computation time for all the patches are selected (by plotting the error rate vs. computation time and selecting the optimum point located within a certain radius range to the origin where the error is low and the computation time is also low). This results in the four most effective and efficient features listed below:

- Energy (Angular Second Moment):

$$ASM = \sum_i \sum_j P(i,j)^2 \tag{4.2}$$

Energy measures the image gray-level distribution and the texture uniformity. ASM is relatively large when the distribution of $P(i,j)$ is more concentrated on the main diagonal.

- Entropy:

$$ENT = -\sum_i \sum_j P(i,j) log P(i,j) \tag{4.3}$$

Entropy measures the disorder of an image. ENT is larger when the value of $P(i,j)$ is more dispersed and it achieves its largest value when all the $P(i,j)$s are equal.

- Correlation:

$$COR = \frac{\sum_i \sum_j (ij)P(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{4.4}$$

Correlation measures the gray tone linear dependencies in an image. $\mu_x, \mu_y, \sigma_x, \sigma_y$ are the means and standard deviations of $P_x(i) = \sum_j P(i,j)$ and $P_y(j) = \sum_i P(i,j)$.

- Homogeneity (Inverse Difference Moment):

$$IDM = \sum_i \sum_j \frac{1}{1+(i+j)^2} P(i,j) \tag{4.5}$$

Homogeneity is inversely proportional to the image contrast feature at constant energy. Smaller gray tone difference in pair elements will contribute to larger value of homogeneity.

The above selected features are considered as the most relevant (or most effective) features, because they have smaller variations across different combinations of displacement and direction parameters, while they capture the information from different texture patterns more effectively. We tested the features on the training image patches and these features gave the best stability and discriminative power compared to the reset of 10 unselected features. We set the distance between the pair of pixels at 4 scales (1, 2, 4, 8) and set the directions at 4 angles (0°, 45°, 90°, 135°). These scale and orientation parameters were examined as the most appropriate setting by applying Chi-square test on the optimal GLCM computed with the selected four features of the training patches. The final GLCM texture feature vector is oh length 64 (4 feature types × 4 direction × 4 distances).

3) SIFT (Scale Invariant Feature Transform) based keypoint feature. SIFT [102] proposed by Lowe is a very popular feature used in computer vision and pattern analysis. SIFT feature has the advantage that it is invariant to changes in scale, rotation, and intensity. The major issues related to extracting this feature include selecting the keypoints and calculating the gradient histogram of pixels in a neighboring rectangular region. In this work, we apply the Difference-of-Gaussians (DoG) operator to extract the keypoints. For each keypoint, the 16×16 pixels in the neighboring region are used. We divide a region into 16 4×4 subregions. For each pixel in a subregion, we calculate the direction and magnitude of its gradient. We quantize the directions into 8 bins, and build a histogram of gradient directions for each subregion. The magnitude of the gradient is used to weight the contribution of a pixel. Finally, the 8-dimensional feature vectors from the eight-bin direction histogram of each subregion are combined and weighted into a 128-dimensional vector to record local information around the keypoint.

4.2.4 SRV Attribute Detector Learning Module

In this module, the SRV attribute detector is trained by using a generative approach based on probability theory. To illustrate the basic idea, consider a scenario in which an image region depicted by an N-dimensional low-level feature vector $\vec{X^N}$ is to be assigned into one of the K SRV attributes $k = 1, ..., K$ in a higher level of semantics. From probability theory we know that the best solution is to achieve the *a posterior probabilities* $p(k|X)$ for a given X and each attribute category k, and assign the attribute with the largest probability score to the region. In the generative model, we model the joint probability distribution $p(k, X)$ of image region features and attributes, and Bayes' theorem provides an alternative to derive $p(k|X)$ from $p(k, X)$:

$$p(k|X) = \frac{p(k, X)}{p(X)} = \frac{p(X|k)p(k)}{\sum_{i=1}^{K} p(X|i)p(i)} \tag{4.6}$$

As the sum in the denominator takes the same value for all the attribute categories, it can be viewed as a normalization factor over all the attributes. Equation (4.6) can be rewritten as:

$$p(k|X) \propto p(k, X) = p(X|k)p(k) \tag{4.7}$$

which means we only need to estimate the attribute prior probabilities $p(k)$ and the likelihood $p(X|k)$ separately. The generative model has the advantage that it can augment the large amount of unlabeled data in a dataset from a small portion of the labeled data.

As defined earlier K denotes the pool of SRV attributes. Let k_i be the *ith* attribute in K. According to the previous section, k_i is assigned to a set of image regions $R_{k_i} = \{r_1^i, r_2^i, ..., r_{n_{k_i}}^i\}$ along with the corresponding feature vectors $X_{k_i} = \{x_1^i, x_2^i, ..., x_{n_{k_i}}^i\}$, where n is the number of regions in an image. We assume the feature vector is sampled from some underlying multi-variate density function $p_X(\cdot|k_i)$. We use a non-parametric kernel-based density estimate [61] for the distribution p_X. Assuming region r_t to be in the test image with feature vector x_t, we estimate $p_X(x_t|k_i)$ by using a Gaussian kernel over the feature vectors X_{k_i}:

$$p_X(x_t|k_i) = \frac{1}{n}\sum_{j=1}^{n}\frac{exp\{-(x_t - x_j)^T\Sigma^{-1}(x_t - x_j)\}}{\sqrt{2^n\pi^n|\Sigma|}} \tag{4.8}$$

Σ is the covariance matrix of the feature vectors in X_{k_i}.

$p(k_i)$ is estimated by using Bayes estimators with a prior beta distribution, the probability distribution of $p(k_i)$ is given by:

$$p(k_i) = \frac{\mu\delta_{k_i,r} + N_{k_i}}{\mu + N_r} \tag{4.9}$$

where μ is the smoothing parameter estimated from the training set, $\delta_{k_i,r} = 1$ if attribute k_i occurs in the training region r and 0 otherwise. N_{k_i} is the number of training regions that contain attribute k_i and N_r is the total number of training regions.

Finally, for each test region with feature vector x_t, the *posterior probability* of observing attribute k_i in K given x_t, $p(k_i|x_t)$ is given by multiplying the estimates of the two distributions:

$$p(k_i|x_t) = (\frac{1}{n}\sum_{j=1}^{n}\frac{exp\{-(x_t - x_j)^T\Sigma^{-1}(x_t - x_j)\}}{\sqrt{2^n\pi^n|\Sigma|}}) \times (\frac{\mu\delta_{k_i,r} + N_{k_i}}{\mu + N_r}) \tag{4.10}$$

For each salient region extracted from a test image I, the occurrence probability of each attribute in that region is inferred by equation (4.10). The probabilities for all attributes are combined into a single vector which is called *region SRV attribute signature.* For a test image with several salient regions, we combine the region SRV attribute signature into a final vector by choosing the max score for each attribute. We name this vector as the *image SRV attribute signature* and it is used as the semantic descriptor for images.

4.2.5 SRV Attribute Co-occurrence Pattern Detection Module

Attribute labels given by human experts as ground-truth semantic descriptions across the entire training image set are used to learn the contextual information based on the attribute label co-occurrences. In this section, we devise a novel approach to discover the co-occurrence patterns of the individual attributes based on network analysis theories. More specifically, we construct an attribute co-occurrence network to record all the pairwise co-occurrence between attributes. The patterns are detected as the communities in a network structure. A similar concept is used in social network to describe a group of people that have tightly-established interpersonal relationships.

4.2.5.1 SRV attribute co-occurrence pattern detection

We first introduce the notion of community structure from the network perspective. One way to understand and analyze the correlations among individual items is to represent them in a graphical network. The nodes in the network corresponds to the individual items (attributes in our case), the edges describe the relationships (attribute co-occurrence in our case), and the edge weights denote the relevant importance of the relationship (co-occurrence frequency in our case).

A very common property of a complex network is known as the community structure, i.e., groups of nodes may have tight internal connections in terms of a large number of internal edges, while they may have less edges connecting each other. These groups of nodes constitute the communities in the network. The existence of community structure reflects underlying dependencies among elements in the target domain. If a group of individual attributes always occur together in the training image set, then an underlying co-occurrence pattern can be defined by these attributes, and this pattern can be used as a priori knowledge in the attribute detection for the test images.

The approach we adopted to detect the communities in the network is modularity optimization [119]. Suppose attributes a_i and a_j in A are represented as two nodes i and j, and suppose i belongs to community C_i and j belongs to community

C_j in a partition. The modularity Q is defined as a qualitative measure of a particular partition on the network in the form of:

$$Q = \frac{1}{2d}\sum_{i,j}[w_{ij} - \frac{w_i w_j}{2d}]\delta(C_i, C_j) \tag{4.11}$$

where d equals to half of the summation of all the edge weights in the network, w_{ij} is the edge weight between i and j, $w_i(w_j)$ equals the summation of the edge weights attached to node $i(j)$, $\delta(C_i, C_j) = 1$ if $C_i = C_j$ and 0 otherwise.

We consider iteratively merging the nodes into communities based on the criterion that the merge of nodes generates a positive modularity gain at each iteration. The modularity gain of moving an outside node i into a community C is evaluated by

$$\Delta Q = [\frac{\Sigma_{in} + k_{i,C}}{2d} - (\frac{\Sigma_{out} + w_{i,C}}{2d})^2] - [\frac{\Sigma_{in}}{2d} - (\frac{\Sigma_{out}}{2d})^2 - (\frac{w_{i,C}}{2d})^2] \tag{4.12}$$

where Σ_{in} represents the sum of edge weights inside C, $w_{i,C}$ equals the sum of weights of edges that link i to C, d is the same as defined in equation (4.7), Σ_{out} is the sum of weights of edges that link outside nodes to nodes in C, w_i is the sum of weights of the edges incident to i. Based on modularity optimization, we propose the following two phase algorithm to detect the attribute co-occurrence patterns in the network:

4.2.5.2 SRV attribute signature refinement with the co-occurrence patterns

The co-occurrence patterns are utilized for refining the detection results on each individual SRV attribute by performing a random walk process [76] over the patterns. We define the distance between two attributes a_i and a_j as

$$D_{a_i,a_j} = \frac{2\ \times\ \#\ of\ CP\{a_i, a_j\}}{\#\ of\ CP\{a_i\}\ +\ \#\ of\ CP\{a_j\}} \tag{4.13}$$

where $\#\ of\ CP\{a_i, a_j\}$ is the number of co-occurrence patterns containing both attribute a_i and a_j. Suppose initially the occurrence probability of attribute a_i in the *image attribute signature* is $s(a_i)$ (given by the generative model), then in the *mth* it-

Algorithm 4: SRV Attribute Co-occurrence Pattern Detection

```
    Input: SRV attribute co-occurrence network.
    Output: Hierarchical SRV attribute co-occurrence patterns.
 1  Partitioning phase:
 2  do
 3      Assign each node a different community tag $C_i, i = 1, ..., N$;
 4      foreach node $i$ in Community $C_i$ do
 5          Remove $i$ from its original community $C_i$;
 6          Add $i$ into each of its neighboring nodes $j$'s community $C_j$;
 7          if $\Delta Q > 0$ computed by (4.12) from placing $i$ to $C_j$ then
 8              Examine the value of $Q_{C_i}$ and $Q_{c_j}$ with $i$ assigned to each neighboring
                community by (4.11);
 9              if $Q_{C_i} \geq 0.3$ && $Q_{c_j} \geq 0.3$ then
10                  Attribute $i$ is shared by the two communities $C_i$ and $C_j$;
11                  Split $i$ into $i$ and $i^{'}$, put them into $C_i$ and $C_j$;
12                  Copy the edges of $i$ incident to other nodes for $i^{'}$;
13              else
14                  Place $i$ into $C_j$;
15          else
16              No node will be moved;
17  while Every node has been traversed && no increase can be achieved for $\Delta Q$;
18  Coarsening phase:
19  foreach Existing community $C_i$ do
20      Replace the entire community $C_i$ by a single node $i$ in the network;
21      Replace the edges between community $C_i$ and its neighboring communities by
        single edges;
22      Compute the weight for a single edge as the sum of old edge weights;
23      Represent internal edges as a self-looped edge with weight equals the sum of
        internal edge weights;
24  Iteration:  Repeat 1 $\rightarrow$ 23 until no positive $\Delta Q$ can be achieved;
```

eration the new value of the probability is formulated by the following random walk process:

$$s_m(a_i) = \alpha \sum_j s_{m-1}(a_j) \cdot D_{a_i,a_j} + (1-\alpha) \cdot s(a_i) \tag{4.14}$$

where α is a weight parameter that takes a value between $(0, 1)$. The above formula can strengthen the occurrence probabilities of the attributes in the same patterns and weaken the isolated ones. The controlling parameter is determined by using the training sets.

4.2.6 Identification Module

The attribute detector learned from the training data is used in the identification module for the testing images. The inputs to the detector are the detected salient regions from the test images as well as the extracted low-level visual features. The output of the detector is the so-called "image SRV attribute signature". The species identification of testing images is performed by comparing testing image signatures with the training image signatures. Therefore, we also build the attribute signatures for the training images. For a training image I, the attribute signature is $S^{|A|}$ with each element $s(a_i) \in \{0, 1\}$ and $s(a_i) = 1$ when image I has regions labeled with attribute a_i and $= 0$ otherwise. We further divide the training images into groups based on their scientific species designation. The element values are averaged across the signatures within each species group for each individual attribute and the obtained signature is called the *species prototype signature.*

The testing image of a species is identified by comparing its image attribute signature with the species prototype signatures of the fifty species. The distance between the two signatures is calculated by the Euclidean distance. The testing image is finally identified as the species with the smallest distance value. If several species have very similar distance values to the testing image, we assign all the species labels to that image, and let the image retrieval system give the final decision on the species based on the feedback from the users who are determined as experts by the retrieval system.

4.2.7 Retrieval & Relevance Feedback Module

We implement a query by example (QBE) paradigm for our retrieval system. QBE is widely used in conventional content-based image retrieval (CBIR) systems when the image meta-data, such as captions, surrounding texts, etc. are not available for keyword based retrieval.

4.2.7.1 Image retrieval using query by example

In the QBE mode, the user is required to submit query in terms of an example specimen image to the system. Finding an appropriate query example, however, is still a challenging problem in the research area of CBIR. In our system, we provide an image browsing function in the user interface, and the user is allowed to browse all the images in the database and submit a query. Images are compared by their content similarity. Each image in the database is represented by a low-level visual feature vector F and a high-level SRV attribute signature S, for a query image Q and a database image Y. The distance between them is calculated by fusing the Euclidean distance over the visual feature vectors and the Earth Mover's distance [138] over the SRV attribute signatures:

$$Dist(Q, Y) = \eta D_{Euc}(F_Q, F_Y) + (1 - \eta) D_{EMD}(S_Q, S_Y) \tag{4.15}$$

where η is the adjusting parameter between the two distance measures and is determined by the long-term cross-session retrieval history working on the subset of training images [184]. If the precision for a particular query is increased when more importance is put on the feature distance, then η is adjusted to a larger value, otherwise it becomes smaller.

Earth Mover's Distance (EMD) is used as a proper measure for comparing signature difference given the pre-defined ground distances for pairs of attributes. The underlying idea of Earth Mover's distance is: given two signatures of attributes, one can be seen as a mass of earth spread in the attribute space, the other as a collection of holes in the same attribute space. EMD is defined as the least amount of work needed to fill the holes with the earth. The ground distance between a pile of earth (an attribute element in the first signature) and a hole (an attribute element in the second signature) corresponds to the amount of work needed to move that pile of earth to the hole (the base metric defined in the attribute space and used to compute the distance between two attributes). In our setting, the ground distance can be obtained by taking the reciprocal of the edge weights between the two attributes in the co-occurrence network which reflects the hardness that two attributes occur together in the images. Let $d(S_Q(a_i), S_D(a_j))$ denote the ground distance between attribute a_i in the query signature and attribute a_j in the database image signature. The Earth Mover's Distance between their signatures is defined as:

$$D_{EMD}(S_Q, S_D) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d(S_Q(a_i), S_D(a_j))}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \tag{4.16}$$

where f_{ij} is called a flow that is transferred from one signature to the other. The EMD is computed by solving all the f_{ij} using linear programming [99]. The EMD can be viewed as a measure of the least amount of work needed to transfer one signature into the other, a unit of work in the process is evaluated by the ground distance.

4.2.7.2 Relevance Feedback

The Relevance feedback (RF) scheme has been verified as a performance booster for our retrieval system. The reason is that RF can capture more information about user's search intention, which can be used to refine the original image descriptors from feature extraction and attribute detection [40].

Our RF approach follows the Query Point Movement (QPM) paradigm as opposed to the Query Expansion (QEX) paradigm. We move the query point in both the feature space and the attribute space toward the center of the user's preference by using both the relevant and irrelevant samples marked by the user at each retrieval iteration. However, before the users' decisions are used to refine the descriptors, their expertise in identifying moth species are evaluated by sample tests when they first enter the system. If an user has 90% accuracy in identifying the species, their relevance feedback will take effect.

Suppose in each retrieval iteration the system returns N images. Let $F = \{f_1, f_2, ..., f_N\}$ denote the visual feature vectors and $S = \{s_1, s_2, ..., s_N\}$ denote the attribute signatures of the retrieved images, and let f_Q and s_Q represents the query descriptors accordingly. The refinement on the descriptors is equivalent to learning projection matrix W_f that transforms $\{f_1, f_2, ..., f_N, f_Q\}$into $\{f_1^{'}, f_2^{'}, ..., f_N^{'}, f_Q^{'}\}$, as well as W_s that transforms $\{s_1, s_2, ..., s_N, s_Q\}$ into $\{s_1^{'}, s_2^{'}, ..., s_N^{'}, s_Q^{'}\}$, by which the query and the relevant images resemble as much as possible in the feature and attribute spaces and deviate from the irrelevant ones.

Let $\mathcal{P}$ and $\mathcal{N}$ denote the sets of positive and negative results. We build pairwise relevant descriptor set Λ_f, Λ_s and pairwise irrelevant descriptor set Ω_f, Ω_s in the following way:

$$\begin{cases} \Lambda_f = \{(f_Q, f_i) | f_i \in \mathcal{P}_f\} \cup \{(f_i, f_j) | f_i, f_j \in \mathcal{P}_f\} \\ \Lambda_s = \{(s_Q, s_i) | s_i \in \mathcal{P}_s\} \cup \{(s_i, s_j) | s_i, s_j \in \mathcal{P}_s\} \\ \Omega_f = \{(f_Q, f_i) | f_i \in \mathcal{N}_f\} \cup \{(f_i, f_j) | (f_i \in \mathcal{P}_f \cap f_j \in \mathcal{N}_f) \cup (f_i \in \mathcal{N}_f \cap f_j \in \mathcal{P}_f)\} \\ \Omega_s = \{(s_Q, s_i) | s_i \in \mathcal{N}_s\} \cup \{(s_i, s_j) | (s_i \in \mathcal{P}_s \cap s_j \in \mathcal{N}_s) \cup (s_i \in \mathcal{N}_s \cap s_j \in \mathcal{P}_s)\} \end{cases} \quad (4.17)$$

After the transformation W_f, the sum of the squared distances of the visual feature pairs in Λ_f is comupted as:

$$\begin{aligned} &\sum_{(f_i,f_j)\in\Lambda_f} (W_f^T f_i - W_f^T f_j)^T (W_f^T f_i - W_f^T f_j) \\ &= \sum_{(f_i,f_j)\in\Lambda_f} Tr[W_f^T (f_i - f_j)(f_i - f_j)^T W_f] \\ &= Tr(W_f^T X_{\Lambda_f} W_f), \end{aligned} \quad (4.18)$$

where $X_{\Lambda_f} = \sum_{(f_i,f_j)\in\Lambda_f} (f_i - f_j)(f_i - f_j)^T$ and Tr is the trace of the matrix. Similarly, we have $Tr(W_s^T X_{\Lambda_s} W_s)$, $Tr(W_f^T X_{\Omega_f} W_f)$ and $Tr(W_s^T X_{\Omega_s} W_s)$. We would like to have the sum of distances from Λ as small as possible and the sum of distances from Ω as large as possible, so have the following objective functions:

$$\begin{cases} \underset{W_f^T W_f = I}{min} Tr(W_f^T X_{\Lambda_f} W_f), \underset{W_f^T W = I}{max} Tr(W_f^T X_{\Omega_f} W_f) \\ \underset{W_s^T W_s = I}{min} Tr(W_s^T X_{\Lambda_s} W_s), \underset{W_s^T W = I}{max} Tr(W_s^T X_{\Omega_s} W_s) \end{cases} \quad (4.19)$$

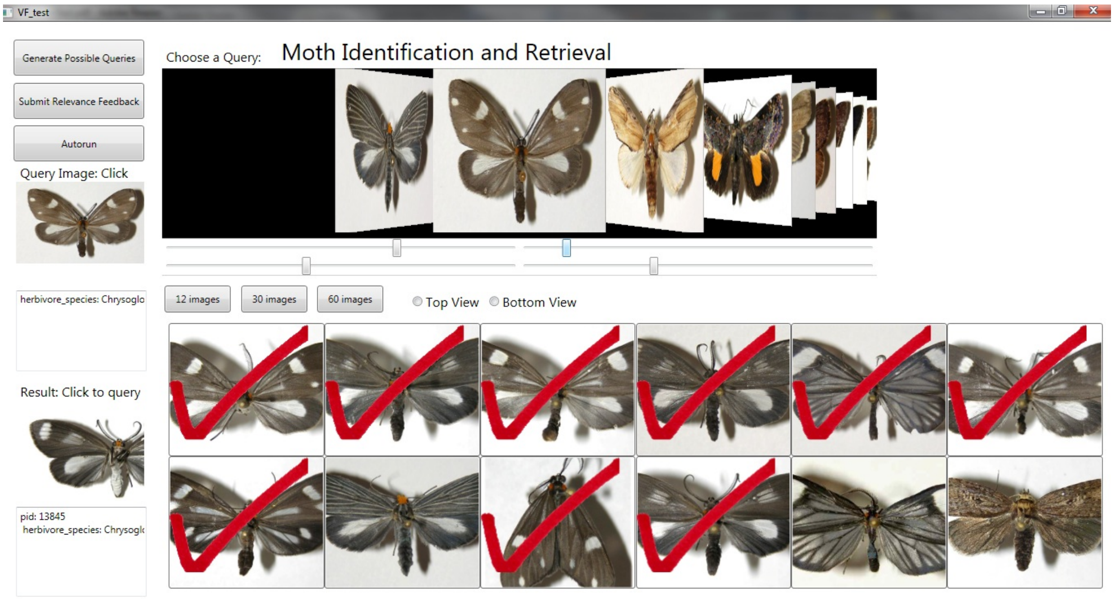
where I is the identity matrix, the purpose of having the constraints $W_f^T W_f = I, W_s^T W_s = I$ is to prevent arbitrary scaling of the projection. The minimization and maximization problems in (5.6) is usually formulated as a *trace ratio* optimization problem [167]:

$$\begin{cases} \underset{W_f^T W_f = I}{max} \frac{Tr(W_f^T X_{\Omega_f} W_f)}{Tr(W_f^T X_{\Lambda_f} W_f)} \\ \underset{W_s^T W_s = I}{max} \frac{Tr(W_s^T X_{\Omega_s} W_s)}{Tr(W_s^T X_{\Lambda_s} W_s)} \end{cases} \quad (4.20)$$

Wang et al. [167] proposed an iterative algorithm to conduct trace ratio optimization, which is adopted in our work to solve the problem in (5.7) and is summarized in Algorithm 2.

Algorithm 5: Trace ratio optimization [167]

```
**Input**: The sum of descriptor distances in the positive and negative sets:
    $X_{\Lambda_f}, X_{\Lambda_s}, X_{\Omega_f}, X_{\Omega_s}$.
**Output**: The transformation matrices $W_f$ and $W_S$
Initialize $W_f^0, W_s^0$ as arbitrary columnly orthogonal matrices such that $(W_f^0)^T W_f^0 = I$ and $(W_s^0)^T W_s^0 = I$.
Set iteration counter $n = 1$.
**repeat**
    Compute $\lambda_f^n, \lambda_s^n$ defined as follows:
        $\begin{cases} \lambda_f^n = \frac{Tr((W_f^{n-1})^T) X_{\Omega_f} W_f^{n-1}}{Tr((W_f^{n-1})^T) X_{\Lambda_f} W_f^{n-1}} \\ \lambda_s^n = \frac{Tr((W_s^{n-1})^T) X_{\Omega_s} W_s^{n-1}}{Tr((W_s^{n-1})^T) X_{\Lambda_s} W_s^{n-1}} \end{cases} \quad (4.21)$
    Solve the following trace difference maximization problem to obtain $W_f^n$ and $W_s^n$ by performing eigen-decomposition of $(X_{\Omega_f} - \lambda_f^n X_{\Lambda_f})$ and $(X_{\Omega_s} - \lambda_s^n X_{\Lambda_s})$:
        $\begin{cases} W_f^n = \underset{W_f^T W_f = I}{argmaxTr}[W_f^T (X_{\Omega_f} - \lambda_f^n X_{\Lambda_f}) W_f] \\ W_s^n = \underset{W_s^T W_s = I}{argmaxTr}[W_s^T (X_{\Omega_s} - \lambda_s^n X_{\Lambda_s}) W_s] \end{cases} \quad (4.22)$
    Set $n = n + 1$.
**until** *convergence*;
**return** $W_f^n$ and $W_s^n$.
```



herbivore_species: Chrysoglo

pid: 13845
herbivore_species: Chrysogl

Figure 4.7: The Screen shot of the system. The images can be browsed in the display window and selected as queries. The "Submit Relevance Feedback" button is used for manual submissions and the "Autorun" button is used for simulated submissions. The species labels are shown in the text area. The user can click to mark the images as relevant, and the rest are used as irrelevant samples automatically. We can show up to 60 retrieved images in dorsal and ventral views.

4.3 Experimental Results

We implemented the system on Microsoft Windows platform using C# ṅet with the Windows Presentation Foundation application development framework. The image database with relevant features and attributes are deployed on MySQL server. The database is set up by importing .txt files with numeric values of the attributes and features, and textual information describing the image properties of the moth images. We show the screenshot of the application in Figure 4.7. We report here the results in two application scenarios: *(i)* moth species identification based on SRV attributes; *(ii)* Moth image retrieval with relevance feedback based on visual features and SRV attributes.

4.3.1 Image Source and System Parameters

Examination of the moth image collection used in this study is introduced in Section 3.1. All 4,530 specimen images used in our experiments were manually labeled with SRV attributes with MBRs by using the tool introduced in Section 4.2.3.2. The species labels are provided by human experts (Dr. Janzen and his colleagues). The labels of the training images are used in the training process. The labels of the testing images are used as ground-truth for validation purposes.

4.3.2 Species Identification Results

We randomly sampled the images into 10 subsets, one subset was held out for testing and the rest of the subsets was used for training the model. This process was repeated ten times by using each subset of images as the testing set. The average of these results on 10 subsets is reported in this paper in Table 4.3. The tuning parameters are summarized in Table 4.2. We evaluated the model performance for each combination of parameters {Q, α, η} on the testing set. We chose the parameter set that maximized the overall performance averaged over the ten testing subsets. The value of the selected parameters are: $Q = 0.3$, $\alpha = 0.6$, $\eta = 0.5$.

4.3.2.1 Evaluation criteria

The performance of the automated species identification is evaluated by the *accuracy* measure. A test image is assigned to the species category for which prototype signature has the smallest distance to the image's SRV attribute signature. The accuracy measure is defined for each species as the number of correctly identified individuals divided by the the total number of specimens of that species in the testing set. A testing image is considered as a correct identification if the species label generated by

Table 4.2: The system parameters for the experiments

Parameter	Section	Setup
Q	Section 4.2.5.1	The value is in the range of [-1,1], we set to 0.3 based on 10 cross-fold validation.
α	Section 4.2.5.2	The value is in the range of [0,1], we set the value to 0.6 based on 10 cross-fold validation.
η	Section 4.2.7.1	The value is in the range of [0,1], the value is set 0.5 based on 10 cross-fold validation.

the program matches with the ground-truth label.

4.3.2.2 Baseline approaches

To demonstrate the effectiveness of our proposed framework for the moth species identification application, we compare with the following approaches as baselines:

- **Baseline-I**: The most basic model that only uses the visual features extracted from Section 4.2.3.4. No SRV attributes and the signature representation has been used. The images are identified purely based on the visual feature vector similarity calculated by using the Euclidean distance.
- **Baseline-II**: Our generative model for individual attribute detection unified with the attribute signature representation serves as the Baseline-II model. However, this model does not include attribute co-occurrence pattern detection and random walk refinement on the SRV attribute signatures.
- **VW-MSI**: We reimplemented a visual words based model based on the available code (http://people.csail.mit.edu/fergus/iccv2005/bagwords.html) online for image classification [150] and name it as "Visual Words based Moth Species Identification" (VW-MSI). We only implemented the appearance model in the approach and ignored the complex spatial structures.
- **SRV-MSI**: Our proposed approach integrated with co-occurrence pattern detection and SRV attribute signature refinement. We name it as "SRV attribute based Moth Species Identification" (SRV-MSI).

We compared the species identification results of the proposed approach with other three approaches in Table 4.3. The best performance as well as the worst performance are made bold in the table. The mean and standard deviation of the accuracy of the experiments conducted for ten times are computed and shown for the fifty species. As we can observe from Table. 4.3, our system performs the best for almost all the fifty species except that VW-MSI outperforms ours in five species: *Neoxeniades luda*, *Isostyla zetila*, *Atarnes sallei*, *Nascus* Burns and *Mysoria ambigua*. This demonstrates the effectiveness of SRV attributes and the co-occurrence patterns used for signature refinement.

The range of the mean identification accuracy of our system on the fifty species is between 0.3455 and 0.7764. The identification accuracy of some of the species is relatively low (e.g. *Hemicephalis agenoria*, *Neoxeniades luda*, *Dasylophia basitincta*, *Dasylophia maxtla* and *Nascus* Burns). When we visually examined the samples from these species, we found that the moth has less discriminative visual patterns or SRV attributes in our scenario on the wings. This phenomenon reflects that our system may lose the power in identifying moth species with dull wings. Specifically, our system achieved low performance in two species categories: *Dasylophia basitincta* and *Dasylophia maxtla*, which have very similar visual appearances. The confusion matrix (we do not show it in the paper for the reason of space limitation) shows that our system mis-identifies the samples from one species into the other species. However, we observe that VW-MSI and other baselines also lose the effectiveness when dealing with moth images with very similar physical appearances. Based on the values of the standard deviation, our system still gives the most stable results across all the species categories compared to the other three approaches.

The total number of SRV attributes manually given to the images by the human experts is 450. As a result, the maximum length of the SRV attribute signature for the images is 450. In order to compare the impact from the vocabulary size of the attributes

Table 4.3: Identification accuracy for the fourty species. The performance of SRV-MSI is greater than all other approaches except for *Neoxeniades luda*, *Isostyla zetila*, *Atarnes sallei* and *Nascus Burns*.

	Baseline I		Baseline II		VW-MSI		SRV-MSI	
Species	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*
Ceroctena amynta	0.2965	0.0321	0.4176	0.0169	0.4318	0.0196	0.4582	0.0174
Eudocima materna	0.4968	0.0257	0.5483	0.0275	0.5764	0.0319	0.5944	0.0209
Eulepidotis folium	0.3910	0.0279	0.4141	0.0264	0.4219	0.0267	0.4482	0.0371
Eulepidotis rectimargo	0.5561	0.0246	0.5875	0.0236	0.5962	0.0233	0.6134	0.0163
Hemicephalis agenoria	0.3314	0.0268	0.3349	0.0302	0.3721	0.0331	0.3931	0.0236
Thysania zenobia	0.4102	0.0327	0.4329	0.0236	0.4623	0.0235	0.4971	0.0356
Chrysoglossa norburyi	0.5472	0.0225	0.5553	0.0253	0.5672	0.0237	0.5752	0.0205
Erbessa albilinea	0.6048	0.0365	0.6324	0.0336	0.6547	0.0136	0.6755	0.0174
Erbessa salvini	0.3562	0.0468	0.3634	0.0425	0.3867	0.0325	0.4143	0.0345
Nebulosa erymas	0.5432	0.0312	0.5647	0.0291	0.5699	0.0257	0.5935	0.0225
Tithraustes noctiluces	0.5438	0.0214	0.5624	0.0331	0.5912	0.0284	0.6086	0.0251
Polypoetes haruspex	0.5247	0.0216	0.5369	0.0234	0.5682	0.0273	0.5906	0.0202
Dioptis longipennis	0.5621	0.0281	0.5746	0.0212	0.5990	0.0187	0.6154	0.0175
Methionopsis ina	0.4721	0.0375	0.4835	0.0367	0.5014	0.0325	0.5102	0.0425
Neoxeniades luda	0.3742	0.0374	0.3852	0.0432	0.4176	0.0396	0.3975	0.0457
Saliana Burns	0.5042	0.0364	0.5356	0.0256	0.5494	0.0275	0.5731	0.0234
Saliana fusta	0.6480	0.0247	0.6597	0.0275	0.6968	0.0214	0.7346	0.0134
Talides Burns	0.5437	0.0256	0.5572	0.0247	0.5854	0.0173	0.6352	0.0176
Vettius conka	0.6417	0.0334	0.6782	0.0148	0.7332	0.0184	0.7544	0.0169
Aroma aroma	0.5437	0.0273	0.6035	0.0245	0.6204	0.0174	0.6461	0.0211
Carystoides escalantei	0.5326	0.0324	0.5487	0.0264	0.5843	0.0222	0.6033	0.0254
Lirimiris guatemalensis	0.3975	0.0421	0.4129	0.0256	0.4615	0.0236	0.4930	0.0249
Isostyla zetila	0.5248	0.0363	0.5392	0.0365	0.5472	0.0251	0.5364	0.0357
Oricia domina	0.4964	0.0368	0.5175	0.0316	0.5389	0.0380	0.5632	0.0195
Scotura leucophleps	0.5014	0.0378	0.5246	0.0217	0.5547	0.0246	0.5757	0.0221
Bardaxima perses	0.4764	0.0371	0.4954	0.0314	0.5327	0.0287	0.5551	0.0307
Dasylophia basitincta	0.3842	0.0457	0.3976	0.0351	0.4029	0.0225	0.4344	0.0275
Dasylophia maxtla	0.3683	0.0416	0.3754	0.0363	0.3938	0.0324	0.4113	0.0278
Nystalea collaris	0.4173	0.0285	0.4326	0.0291	0.4852	0.0257	0.5021	0.0274
Tachuda discreta	0.3647	0.0321	0.4056	0.0249	0.4303	0.0352	0.4512	0.0269
Atarnes sallei	0.6084	0.0372	0.6396	0.0278	0.7174	0.0147	0.7059	0.0187
Dyscophellus phraxanor	0.5483	0.0364	0.5731	0.0381	0.6295	0.0331	0.6494	0.0362
Tithraustes lambertae	0.6053	0.0271	0.6056	0.0374	0.6324	0.0289	0.6713	0.0285
Entheus matho	0.6153	0.0490	0.6273	0.0411	0.6308	0.0271	0.6534	0.0279
Hyalothyrus neleus	0.6472	0.0394	0.6717	0.0285	0.6954	0.0192	0.7106	0.0168
Nascus Burns	0.3258	0.0173	0.3394	0.0314	0.3547	0.0390	**0.3455**	0.0372
Phocides nigrescens	0.6138	0.0442	0.6359	0.0321	0.6784	0.0179	0.6797	0.0171
Quadrus contubernalis	0.6432	0.0316	0.6572	0.0257	0.6933	0.0271	0.7096	0.0263
Urbanus belli	0.5276	0.0164	0.5713	0.0268	0.5944	0.0196	0.6132	0.0254
Melanopyge Burns	0.6261	0.0255	0.6527	0.0275	0.6798	0.0138	0.6930	0.0214

Table 4.4: The ranges of accuracy as a function of the number of SRV attributes and the number of visual words used in the experiments. The bold number indicates the largest accuracy for each approach.

	Accuracy range			
	VW-MSI		SRV-MSI	
Number of attributes/visual words	*Lower bound*	*Upper bound*	*Lower bound*	*Upper bound*
50	0.2137	0.4282	0.2542	0.4673
100	0.2563	0.4356	0.2716	0.4927
150	0.2918	0.4847	0.3164	0.5574
200	0.3157	0.6658	0.3374	0.6923
250	0.3578	**0.7271**	0.3763	0.7567
300	0.3334	0.7016	0.3431	**0.7764**
350	0.3126	0.6567	0.3267	0.7637
400	0.2876	0.6145	0.3178	0.7521
450	0.2747	0.5983	0.3027	0.7278

and the visual words for VW-MSI and SRV-MSI, we set the maximum size of the visual words vocabulary to 450 as well. The SRV attributes and the visual words are ranked in the relative vocabulary based on the number of appearance in the image collection.

We change the number of attributes and visual words used in the experiments and show the corresponding accuracy variation in Table 4.4. The best performance achieved by our approach is marked bold in the table. We can observe that our system achieves the best performance when we use approximately 300 SRV attributes. For the visual words based approach, the best performance is achieved when the number of visual words is around 250. It is obvious that the accuracy of VW-MSI drops very fast and has a large range when the number of visual words exceeds 300. However, our SRV attribute based approach has relatively small variations across different attribute settings which demonstrates that it is less sensitive to the vocabulary size compared to VW-MSI.

4.3.3 Image Retrieval Results

To test the performance of our SRV attribute based approach for image retrieval with the proposed relevance feedback scheme, like for species identification in Section 4.2, we divided the entire image dataset into 10 folds. The parameters are determined using the same scheme as described in Section 4.2.1. We set the number of attributes to 300. In order to reduce the amount of work of submitting relevance feedback that are required by users, we propose to simulate the user interaction by launching queries and submitting feedback automatically by the system. The simulated process works in the following way: the system compares the ground-truth species labels of the retrieved images with the query, if the species label matches the query, the system will mark the image as relevant, otherwise, the image is marked as irrelevant. By doing this, we assume the relevance feedback provided by the users will always by correct (i.e., users will only mark the relevant images as those from the same species category as the query). Note that the ground-truth is only used by the system to judge the relevance of the retrieved images. It is not involved in comparing image similarity in the retrieval procedure. For each query, we request the users or the system to provide five iterations of relevance feedback. We have half of the queries in each species category launched by the users and the other half simulated by the system. The results are computed based on the combination of the two methods.

4.3.3.1 Evaluation criteria

In each iteration, the retrieval precision is evaluated by the rank of the relevant images. Further statistical evaluation of the averaged precision for each species relies on standard image retrieval measure: *Mean average precision of top D retrieved images* over all the query images from a specific species category. Let D be the number of retrieved images and R be the relevant ones with size $|R|$. Given a query Q, the average precision is defined as $AP(Q) = \frac{1}{|R|} \sum_{i=1}^{|R|} \frac{i}{Rank(R_i)}$, and the mean average precision (MAP) is the averaged AP over all the testing images.

4.3.3.2 Baseline approaches

To demonstrate the effectiveness of our proposed retrieval framework, we use the following approaches as the baselines to compare the results:

- **Baseline-I**: The proposed image retrieval framework without relevance feedback scheme.

- **Baseline-II**: We reimplemented an insect image identification approach [169] and integrated it into our retrieval framework with five iterations of relevance feedback process. The features used are a combination of color, shape and texture features and there is no higher level image descriptor like our SRV attribute that has been used in the original approach.

- **SRV-IR**: Our proposed retrieval framework with relevance feedback scheme based on the SRV attributes.

We show the top twelve retrieved images in the application interface. However, the application can be adjusted to show more images upon request. Table 4.5 summarizes the mean averaged precision from the three approaches for all the fifty species. As we can observe, when RF scheme is applied (Baseline-II and SRV-IR), the mean averaged precision is increased compared to the retrieval without RF (Baseline-I), which demonstrates the effect of human interaction in improving the retrieval performance. When more retrieval iterations are involved in the searching process, and when more iterations of relevance feedback are provided, the system can find more relevant images matching user's search intention. In the two approaches that adopts relevance feedback scheme, our approach which uses SRV attribute based image descriptor outperforms Baseline-II for all the species categories. The system response time for each individual query for a database of 1000 images is around 150ms. For a database of 4000 images the response time for each individual query is approximately 500ms.

4.4 Conclusions

In summary, this paper has reported a novel insect species identification and retrieval system based on wing attributes in the moth image dataset. The purpose of the research is to design computer vision and pattern recognition approaches to conduct automated image analysis that can be used by the entomologists for insect studies. We have demonstrated the effectiveness of our system in species identification and image retrieval for fifty moth species.

The dataset we used contains around 4,530 images which could be easily extended to larger sizes in the future to test the scalability of the system. Overall, our system achieves the best performance compared to the baseline approaches in identification and retrieval. The identification accuracy is over 70% on the image collection and the mean average precision reaches 70% as well for some of the species.

A significant diffference between our work and the similar ones in insect identification is that, we provide an intermediate-level feature, namely, the SRV attributes, which function as a bridge, to narrow the semantic gap between machine understanding and human interpretation of the images. We are excited to see that SRV attributes successfully capture the visual patterns on the moth wings at a higher semantic level

Table 4.5: Comparison of the retrieval performance for the fifty species.

Mean Average Precision							
Species	BL-I	BL-II	SRV-IR	*Species*	BL-I	BL-II	SRV-IR
Ceroctenaamynta	0.4096	0.3872	0.4571	*Bardaximaperses*	0.2836	0.3064	0.3275
Eudocimamaterna	0.4538	0.4170	0.4764	*Dasylophiabasitincta*	0.3538	0.4152	0.4658
Eulepidotisfolium	0.3824	0.4115	0.4745	*Dasylophiamaxtla*	0.3628	0.3738	0.4145
Eulepidotisrectimargo	0.5572	0.5069	0.6130	*Nystaleacollaris*	0.3427	0.3735	0.3841
Hemicephalisagenoria	0.4187	0.3950	0.4712	*Tachudadiscreta*	0.2917	0.2978	0.3114
Thysaniazenobia	0.4104	0.3933	0.4705	*Atarnessallei*	0.5832	0.6224	0.6778
Chrysoglossanorburyi	0.5856	0.5710	0.6786	*Dyscophellusphraxanor*	0.5324	0.5799	0.6128
Erbessaalbilinea	0.6045	0.5972	0.7153	*Tithrausteslambertae*	0.4846	0.4472	0.5315
Erbessasalvini	0.4587	0.4311	0.5478	*Entheusmatho*	0.4796	0.4925	0.5486
Nebulosaerymas	0.5219	0.5346	0.5857	*Hyalothyrusneleus*	0.6042	0.6584	0.6971
Tithraustesnoctiluces	0.5486	0.5148	0.5749	*NascusBurns*	0.2396	0.2846	0.3167
Polypoetesharuspex	0.5745	0.5237	0.5964	*Phocidesnigrescens*	0.5755	0.5942	0.6398
Dioptislongipennis	0.4816	0.4754	0.5048	*Quadruscontubernalis*	0.6492	0.7047	0.7168
Methionopsisina	0.3581	0.3847	0.3994	*Urbanusbelli*	0.5693	0.5480	0.5724
Neoxeniadesluda	0.3625	0.3827	0.4117	*MelanopygeBurns*	0.6454	0.6845	0.6992
SalianaBurns	0.5317	0.5485	0.5884	*Myscelusbelti*	0.6715	0.7047	**0.7673**
Salianafusta	0.6046	0.5917	0.6459	*Mysoriaambigua*	0.4917	0.4802	0.5746
TalidesBurns	0.5154	0.5308	0.5742	*Dicentriarustica*	0.3969	0.4105	0.4453
Vettiusconka	0.6296	0.6115	0.7135	*Farigiasagana*	0.2946	0.3072	0.3418
Aromaaroma	0.4537	0.4425	0.5289	*Hapigiodessigifredoma*	0.3634	0.3728	0.4051
Carystoidesescalantei	0.5046	0.4672	0.5274	*Malocampamatralis*	0.4746	0.4869	0.5537
Lirimirisguatemalensis	0.3234	0.3456	0.3753	*MeragisaJanzen*	0.5643	0.5756	0.6683
Isostylazetila	0.5924	0.5483	0.6175	*Naprepahoula*	0.3748	0.4245	0.4886
Oriciadomina	0.4641	0.4547	0.5044	*Pseudodryaspistacina*	0.2975	0.3174	0.3531
Scoturaleucophleps	0.5179	0.5357	0.5678	*Rifargiadissepta*	0.5648	0.5247	0.6190

and generate better results consequently.

However, the discriminative power of our system drops when the moth species contain highly similar visual properties. This could cause potential failure in both identification and retrieval once more images are included in the dataset that belong to different species categories, however, share strong visual patterns on the wings. These cases would be difficult for humans as well.

Future research will include investigations on more effective feature and attributes as well as more advanced learning approaches which could address both the scalability and discrimination issues.

Chapter 5

Understanding Dynamic Social Grouping Behaviors of Pedestrians

5.1 Introduction

Consider a video clip recording a number of pedestrians walking in an outdoor (indoor) environment such as a square (hall). Imagine an algorithm that is able to analyze the video and answer the questions like: Are these people evacuating from an emergent situation? Are they gathering for a special event? By just looking at each individual it could be very hard to train the computers to understand these high-level concepts from the low-level visual representations. In this paper we introduce a new model for analyzing social behaviors among pedestrians: rather than treating each person in isolation, we analyze their social grouping behaviors so as to reinforce the recognition of movements of each individual in a group. Our approach is inspired by recent achievements in computer vision and pattern recognition where the correlations of semantic or geometrical concepts are utilized as extra contextual information for recognizing objects in complex scenes [57]. In our work, pedestrian detection and interactions are enforced by taking the advantage of contextual information that comes from within-group positional, velocity and directional distance consistences. This provides our approach the robustness to pedestrian walking behavior analysis from dynamic cluttered background, occlusions among pedestrians, illumination and viewpoint changes, or the variations of backgrounds caused by mobile cameras such as smart-phones.

It is important to understand the collective social behaviors at a group level in many real-world scenarios. For example, people tend to participate or leave an event with herding behavior [114]. When crowd of people evacuate from an emergent situation, they leave with the members in their original group [124], the direction of the group is usually determined by the fastest member and the speed of the group is limited by the slowest member [65]. Computer vision techniques, such as multi-people tracking in crowded scenes [45, 197], crowd segmentation [162] have made tremendous progress in recent years and they provide the opportunities to solve real-world challenging problems such as recognition of human behaviors at the activity and event level that far exceeds the conventional capabilities of a surveillance system.

In this paper, we attempt to achieve a higher level understanding of crowd behaviors in terms of social groups and interaction patterns that are displayed while they are traveling together. A social group of pedestrians consists of people with shared walking patterns such as change of directions, change of speeds, avoiding obstacles, etc [24]. In particular, we explicitly explore the dynamic properties of social groups that capture the spatio-temporal changes such as splitting and merging of people. Determing the dynamic group structure of a crowd provides the basis for further high-level analysis

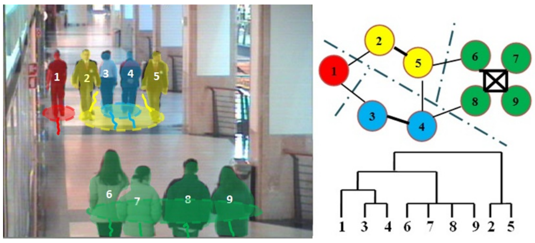


Figure 5.1: Left: A real-world video frame (from CAVIAR dataset) shows that people are walking in groups. Individuals and related trajectories are labeled with numbers and the potential social groups among them are marked in different colors. Right: A snapshot restored from evolving tracklet interaction network (ETIN) representation at a given time interval (top) and a hierarchical social group structure discovered by the proposed approach (bottom).

of events involving social interactions within and across groups.

We propose to detect the social groups of pedestrians based upon the state-of-the-art pedestrian detector and reliable tracklet generation techniques. Our main contribution, is that we explore the evolving social group property among tracklets in a network structure, which we call "*evolving tracklet interaction network*" (ETIN). Based on the social psychological models of collective behavior, the reliable tracklets generated from detection responses are represented as nodes in ETIN with incident edges indicating the social interactions and grouping behaviors (see Figure 5.1). The significance of social grouping behavior between nodes is defined by the edge weights. Tracklets from pedestrians in a potential group will have denser spatio-temporal co-occurrences reflected by larger edge weights in ETIN compared to the tracklets from the pedestrians outside the group. We also propose to address the dynamic changes of social groups in ETIN explicitly which is similar to detecting evolving communities that exist in many common social networks such as Facebook and Twitter.

5.1.1 Contributions of This Chapter

We validate our framework extensively on multiple video datasets that are collected from indoor/outdoor public scenes with elevated viewpoints which is the typical setting of surveillance cameras. We compare the results from our group understanding algorithms with manually labeled ground-truth group IDs in a quantitative way. Our work builds upon the recently proposed techniques in the literature on tracking by detection responses and tracklet association [16, 26, 78, 128, 187, 193]. **Our contributions** are four-fold:

1. We propose a novel evolving tracklet interaction network (ETIN) to depict social grouping behaviors of pedestrians from reliably built tracklets of individuals which embody meaningful spatio-temporal interactions of individuals.

2. We explicitly explore the dynamic property of social groups by providing adaptation schemes for nodes and edges in ETIN representation. Our approach has not only the power of updating the network of tracklets in a very efficient manner, but also has the ability to trace the evolution of the network over time.

3. We introduce a novel modularity optimization based group detection algorithm that detects the *hierarchical* social group structure with a distance metric reflecting the spatio-temporal interactions among the pedestrians. We also provide a unified framework that addresses social group detection refinement and pedestrian tracklet association in an iterative manner.

4. Experimental results and comparison with current techniques using several datasets show that our approach is robust in medium crowd-density scenarios. We find agreement between the predicted social groups and the human-understanding of

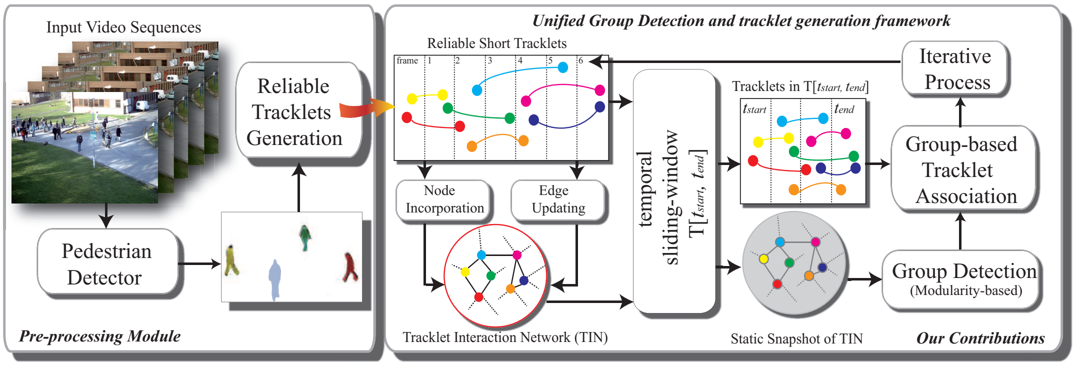

Input Video Sequences
Reliable Tracklets Generation
Pedestrian Detector
Pre-processing Module
Unified Group Detection and tracklet generation framework
Reliable Short Tracklets
frame
1
2
3
4
5
6
Node Incorporation
Edge Updating
Tracklet Interaction Network (TIN)
temporal sliding-window T[t_{start}, t_{end}]
Tracklets in T[t_{start}, t_{end}]
t_{start}
t_{end}
Static Snapshot of TIN
Iterative Process
Group-based Tracklet Association
Group Detection (Modularity-based)
Our Contributions

Figure 5.2: The block diagram of our evolving tracklet interaction network (ETIN) framework for understanding social grouping behavior of pedestrians.

the group structures.

Our work in this paper provides a novel way for social grouping behavior understanding by representing tracklets of pedestrians and their correlations in a network structure which is original in the field. We also provide a framework that iteratively refines the pedestrian tracklet association and social group detection. Our method differs in three ways from the related work of social group recognition: (1) We detect groups in different sizes. In addition, our detected groups are generated in a hierarchical form where groups are captured at different granularities. (2) Our model explicitly handles dynamic changes of social groups, i.e., merging and splitting, in an effective and efficient manner. (3) Our model is built upon tracking by detection techniques where reliable short-term trajectories, or tracklets are available. The social grouping behaviors are, therefore, captured for a period of time in a consistent manner.

5.2 Technical Approach

As illustrated in Figure 5.2, the main focus of this work is to understand the dynamic social grouping behavior of pedestrians by using surveillance videos and developing techniques which provide an automated way to quantitatively analyze videos instead of spending hundreds of person hours to watch and manually labeling them. We name our approach the Evolving Tracklet Interaction Network (ETIN) based dynamic social grouping behavior analysis.

The walking behaviors of pedestrians are represented by their trajectories in the frames. However, it is often a non-trivial task to acquire reasonable trajectories in an automated way for pedestrians in a crowded or semi-crowded environment, because of the occlusions among pedestrians. In this regard, it becomes necessary to track people in a given video for a few seconds without occlusions and yield short-term trajectories, called *tracklets*, and hypothesize pedestrian groups based on these reliable tracklets. The next step is to merge and link these tracklets into long-term trajectories using the detected social groups as contextual information. The hypothesis is that pedestrians in the same group should have very similar trajectories. If some of the trajectories are broken because of occlusion, the rest of the trajectories that are complete in the same group can place useful constraints on associating the fragments. This step plays a critical role in accurately detecting long-term groups and their dynamic changes in the future. We, therefore, provide an unified framework that iteratively discovers social groups from reliable tracklets and identify stable and coherent trajectories of pedestrians that benefits from the group contexts.

We represent the interactions among tracklets by using the proposed *evolving tracklet interaction network* (ETIN) and detect social groups using the modularity-based algorithms. Each tracklet is initialized as a node with corresponding information such as the starting and ending frames of the tracklet that is incorporated into ETIN. The relationship between existing nodes and the new node is measured by the edge weights based on the spatio-temporal interactions of the tracklets. Existing edges also need to be updated each time a new node is incorporated because of the transitive property of social grouping behavior, i.e., the social interactions between two existing nodes should be strengthened when a new node is appended with strong connections to both nodes. In order to reduce the time complexity of updating edge weights, we propose an efficient algorithm that takes advantage of the prior social group information and update the edges in an accelerated way.

To study the dynamic property of social groups such as formation, termination, splitting and merging, it is essential to characterize the transitions that go through a network at different time instants along the video. For this purpose, we utilize temporal snapshots to review static versions of the evolving network at different time intervals by applying time sliding windows in the network. In each snapshot, the nodes are kept that have some temporal overlaps with the time sliding window with corresponding edges. The social groups are then detected from the static ETIN for this specific time interval. This is formulated as a community detection problem and solved by modularity optimization that maximizes the within-group connections and minimizes the between-group connections. In the following we describe major components of the system shown in Figure 2.

5.2.1 Preprocessing Module

We detect pedestrians in each frame using pre-trained deformable part-based detector [56]. In order to lower the percentage of false positives, we explicitly tune the detector to exclude partially occluded people. We also remove detection responses that are of inappropriate sizes as judged by camera calibration. The detections are chained together in a dual-threshold/conflicting pairs data association step to generate short-term tracklets [78]. The output is a set of tracklets that eliminate identity switches.

5.2.2 Evolving Tracklet Interaction Network

For each tracklet x from the output of above procedure, we record the attributes in the format of $x(ID,\ tuple\ set\ \{c_{t_i}, v_{t_i}\}, t_i \in [t_{start}, t_{end}])$, where ID is an unique number used as the index of the tracklet, t_{start}, t_{end} are the corresponding starting and ending frames, tuple $\{c_{t_i}, v_{t_i}\}$ records the centroid of detection c projected onto the ground plane and the estimated velocity vector v at a given time instant (frame) t_i. We initialize nodes and incorporate them into TIN for the tracklets in the order of their t_{start} attribute. Each node is also assigned with the corresponding tracklet's attributes. The interactions between individual nodes are modeled as pairwise spatio-temporal co-occurrences and we represent them as edges in the network. Edge weight indicates the significance of a specific interaction. For a given pair of tracklets, we categorize their interaction into two types based on whether they have a temporal overlap: 1) interaction of tracklets with overlap and 2) interaction of tracklets without overlap.

For the *first* type of interaction, we define the temporal overlap as $\Gamma = [t_0, t_1]$ of length $(t_1 - t_0 + 1)$ frames. The interaction between two tracklets is measured by the weighted sum of aggregated positional, velocity and directional distances. Given two tracklets x_i and x_j, the distances are defined as:

$$\begin{cases} D^p(x_i, x_j) & = & 1 - exp(-\frac{\sum_{t=t_0}^{t_1} ||c_i^t - c_j^t||}{|\Gamma|\rho^p}) \\ D^v(x_i, x_j) & = & 1 - exp(-\frac{\sum_{t=t_0}^{t_1} ||v_i^t - v_j^t||}{|\Gamma|\rho^v}) \\ D^d(x_i, x_j) & = & 1 - exp(-\frac{c_i^{t_1} - c_i^{t_0}}{||c_i^{t_1} - c_i^{t_0}||} \cdot \frac{c_j^{t_1} - c_j^{t_0}}{||c_j^{t_1} - c_j^{t_0}||}) \end{cases} \tag{5.1}$$

where ρ^p and ρ^v are scaling factors for tuning the aggregated distance. The double vertical bar $(|| \cdot ||)$ represents the $L2$ norm of a vector. All the three distance measures are scaled into the range $[0, 1]$ by exponential normalization. Aggregating the distances over time increases the robustness for capturing dynamic social grouping be-

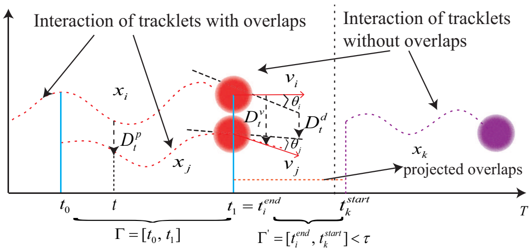


Figure 5.3: Two types of tracklet interactions are shown in the left and right side. The two tracklets with overlapped interaction are marked in red and the other tracklet without overlap is marked in purple. The importance of the interaction is either calculated based on their positional, velocity and directional distances based on the temporal overlapping interval or the distances based on the projected overlapping interval.

haviors. Tracklets that are closer to each other and have similar velocities and directions for a longer time will yield smaller distances. The final pairwise interaction is defined as:

$$e_{ij}^{\Gamma} = exp(-(\omega_1 \cdot D^p + \omega_2 \cdot D^v + (1 - \omega_1 - \omega_2) \cdot D^d)) \tag{5.2}$$

where ω_1 and ω_2 are the weights to adjust the importance of each factor. We use equal weights in our setting to combine the three distance measure into a final tracklet interaction importance measure that is computed over the temporal interval of overlap.

For non-overlapping tracklets x_i and x_j, suppose $t_i^{end} < t_j^{start}$ and the time interval $\Gamma = [t_i^{end}, t_j^{start}] < \tau$ where τ is a threshold, we determine the potential spatio-temporal interaction between them in the projected overlap interval $[t_i^{end}, t_j^{start}]$ based on the motion model. Let $t \in [t_i^{end}, t_j^{start}]$, we estimate the centroids of both tracklets at frame t by Eq. (5.3).

$$\begin{cases} c_i^t &= c_i^{t_i^{end}} + v_i^{t_i^{end}} \cdot (t - t_i^{end}) \\ c_j^t &= c_j^{t_j^{start}} + v_j^{t_j^{start}} \cdot (t_j^{start} - t) \end{cases} \tag{5.3}$$

The velocities are assumed to be constant in the interval and represented by $v_i^{t_i^{end}}$ and

$v_j^{t_j^{start}}$. We compute the interaction importance for the *second* type of interaction by replacing the parameters in Eq. (5.1) with $c_i^t, c_j^t, v_i^{t_i^{end}}, v_j^{t_j^{start}}$ and $\Gamma = [t_i^{end}, t_j^{start}]$, and repeat Eq. (5.2). Finally, the computed values from Eq. (5.2) are used as the edge weights between pairs of nodes representing the tracklets in ETIN. The two types of tracklet interaction and the distances are illustrated in Figure 5.3. The interaction importance is used as the edge weights when connecting two nodes representing the tracklets in the ETIN.

For each new node, respective edges are added based on the conditions $t_{new}^{start} < t_{existing}^{end} + \tau$ for a non-negative threshold τ. However, the edge weights between existing nodes also need to be updated because of the social group transitivity. For example, two existing nodes x_i, x_j initially have a small interaction degree. When a new node x_k is added, both e_{ik} and e_{jk} are large which implies a high probability that x_i and x_k are in a social group, so are x_j and x_k. And if $\frac{1}{e_{ij}} \geq \frac{1}{e_{ik}} + \frac{1}{e_{jk}}$, in this case, x_i, x_j, x_k should be in a same group and e_{ij} also needs to be modified accordingly.

Consider N existing nodes $\{x_1, x_2, ..., x_n\}$ and a new node x_k, we can calculate $e_{ik}, i \in \{1, ..., n\}$ for any pair of nodes (x_i, x_n), and compare if $\frac{1}{e_{ij}} \geq \frac{1}{e_{ik}} + \frac{1}{e_{jk}}, i, j \in \{1, ..., n\}$ & $i \neq j$. However, if the number of nodes in the network is large, the computation will take a lot of time. In order to reduce the computational cost, we propose a group detection based node incorporation and edge updating scheme as illustrated in Figure 5.4.

First, we denote the constructed ETIN at current frame t as G_t. We detect the groups of nodes using the approach proposed in Section 5.2.3 and the groups are represented as $\{g_1, g_2, ..., g_m\}$. Further, we compute the intergroup closeness between any pair of groups by the symmetric Hausdorff similarity measure $H(g_i, g_j) = \frac{h(g_i,g_j)+h(g_j,g_i)}{2}$

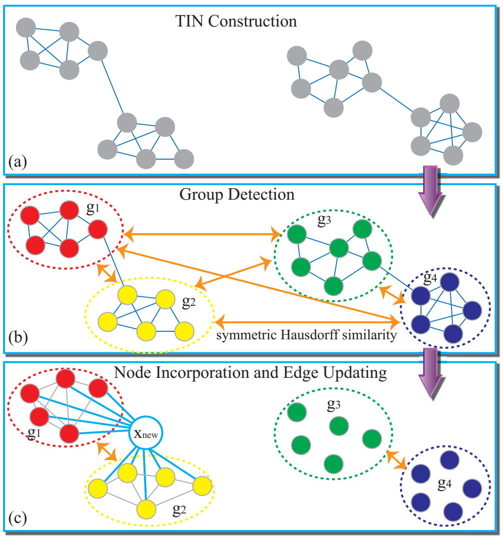

TIN Construction
(a)
Group Detection
g1
g2
g3
g4
symmetric Hausdorff similarity
(b)
Node Incorporation and Edge Updating
Xnew
g1
g2
g3
g4
(c)

Figure 5.4: The new node incorporation and edge updating scheme for the evolving ETIN. (a) The original ETIN. (b) Detection the social groups among nodes based on the modularity optimization. The symmetric Hausdorff similarity is calculated for each pair of groups. (3) When a new node x_{new} is added, the interactions to other nodes are computed only for the nodes in the groups that have distances to x_{new} below a certain threshold.

where $h(\cdot, \cdot)$ is defined by Eq. (5.4).

$$h(g_i, g_j) = \frac{\sum_{i=1}^{|g_i|} \cdot \sum_{k=1,g_j}^{\lceil |g_j|/2 \rceil} sort(e_{ij})_k}{|g_i| \times \lceil |g_j|/2 \rceil} \tag{5.4}$$

where the sort function arrange e_{ij} in descending order and we use the top-k, k equals half size of the second group. Hausdorff metric is popular in computing the similarity among nodes in two finite sets. g_i and g_j are considered to be close to each other if every member in g_i has large interaction importance to at least half of the members in g_j. The idea is similar to the concept of group expansion introduced in [108].

Using Hausdorff criterion, we set up an appropriate threshold ϵ, and two groups g_i, g_j are considered as neighbouring groups if $H(g_i, g_j) > \epsilon$. When a new node comes in, we compare its averaged interaction degree e_g^{avg} to a group g with another properly chosen threshold ϵ' to see if $e_g^{avg} = \frac{1}{|g|}\sum_{i=1}^{|g|} e_{i \cdot new} > \epsilon'$ which indicates the new node is interacting with the group g. We chose $\epsilon' >> \epsilon$ so that any non-neighbouring groups of

g according to ϵ are not interacting with the new node. In this way, we only need to compute the interactions to g and its neighbouring groups and update the edge weights in these groups. For the example presented in Figure 5.4(c), $e_{g1}^{avg} > \epsilon'$, we calculate the interactions for x_{new} and nodes in g_1 as well as the nodes in the neighbouring group g_2 and updating the corresponding edges and avoid the calculation for non-neighbouring groups g_3 and g_4. The entire process is summarized in Algorithm 5.1.

5.2.3 Social Group Detection

We make use of the temporal snapshots to examine static versions of ETIN at different time intervals. We detect the social groups from a restored static ETIN in a given temporal window using *modularity* measure [117].

Definition: Let $G = (V, E)$ denote a varying tracklet interaction network where V represents unique tracklets and E the interactions that exist among the tracklets. We define a temporal snapshot $S_i(V_i, E_i)$ of G to be a network representing only tracklets and interactions active in a particular time interval $[t_i^{start}, t_i^{end}]$, called the snapshot interval.

A social group, in our case, is defined as a group of nodes in a specific snapshot that has large internal interaction importance. On the other side, nodes in the group will have weak interactions to the outside nodes. A common way towards detecting communities of people based on the links in a social network is to recursively divide the entities in the complete network into subgroups. We naturally transform the group analysis into finding a method from the social network perspective. In order to quantify the goodness of a network partition, *modularity* has been widely accepted as a measurement of the partition which has been found to be robust and effective in many real world networks [117].

Algorithm 6: New node incorporation and edge updating

```
Input: Current ETIN, $\{x_1, ...x_n\}$, $\{e_{ij}\}, i, j \in \{1, ..., n\}, x_{new}$
Output: Evolved ETIN with $x_{new}$ incorporated and edges updated
Step one: Detect social groups $G = \{g_1, ..., g_m\}$ in ETIN by the approach
proposed in Section 5.2.3 ;                /* Group distance calculation.  */
foreach each $g_i$ in $G$ do
    foreach each $g_j$ in $G$ and $j \neq i$ do
        Calculate the inter-group closeness by $H(g_i, g_j)$;
        if $H(g_i, g_j) > \epsilon$ & $g_j \notin g_i^{neighbor}$ then
            Add $g_j$ in $g_i^{neighbor}$;
        else if $g_j \notin g_i^{non-neighbor}$ then
            Add $g_j$ in $g_i^{non-neighbor}$;
Copy $G$ to $G'$ as $x_{new}$'s candidate group set;
do
    foreach $g_i$ in $G'$ do
        if $e_{g_i}^{avg} = \frac{1}{|g_i|} \sum_{j=1}^{|g|} e_{j \cdot new} > \epsilon'$ then
            Updating $e_{j \cdot new}$ where $x_j \in g_i$, delete $g_i$ from $G'$;
            Updating $e_{k \cdot new}$ where $x_k \in g_i^{neighbor}$;
            Delete $g_i^{neighbor}, g_i^{non-neighbor}$ from $G'$;
            if $\frac{1}{e_{jk}} \geq \frac{1}{e_{j \cdot new}} + \frac{1}{e_{k \cdot new}}$ where $x_j, x_k \in g_i || g_i^{neighbor}$ then
                Update $e_{jk}$ by $max(e_{j \cdot new}, e_{k \cdot new})$;
while $G' \neq \varnothing$;
return Updated ETIN;
```

Basically, modularity is the fraction of connections within groups subtracting the expected links of the same quantity of node degrees while the connections are distributed in a random way. Usually, larger value of modularity indicates more significant social grouping phenomenon of nodes. Therefore, our goal is to divide the nodes into groups such that the modularity of the entire network is maximized.

Problem Definition: Given the evolving $G = (G_0, G_1, ..., G_n)$ where G_0 is the snapshot at the first snapshot interval, and the rest Gs are the snapshots obtained by $(G_0 + i * \Delta G)$. The problem is to find an adaptive algorithm that efficiently identify the groups at any snapshot interval utilizing the information from the previous interval.

The modularity Q_{ij} of two nodes x_i, x_j measures the difference between their connection strength and expectation of random pair of nodes in the current snapshot of ETIN. Suppose the neighboring node set of node x_i is N_i where each node is connected by an edge to x_i, the modularity Q_{ij} is defined as,

$$Q_{ij} = e_{ij} - \frac{\sum_{k \in N_i} e_{ik} \cdot \sum_{k \in N_j} e_{jk}}{\sum_{x_m, x_n \in TIN} e_{mn}} \tag{5.5}$$

Initially, we assign all the nodes in one group, the modularity Q of the entire network is the summation of the Q_{ij}s of any pair of nodes. However, if we divide the nodes into two groups, we use a label vector $s \in \mathbb{R}^n$ to denote the group of each node. If an element $s_i = +1$, the corresponding node is assigned to the first group, and $s_i = -1$ otherwise, and the modularity of the network changes to:

$$Q' = s^T \cdot Q \cdot s = s^T \cdot \sum_{i,j} [e_{ij} - \frac{\sum_{k \in N_i} e_{ik} \cdot \sum_{k \in N_j} e_{jk}}{\sum_{m,n} e_{mn}}] \cdot s \tag{5.6}$$

The element values in the vector s are determined by first representing Q in the matrix format, eigen-decomposing it into eigenvalues and eigenvectors, and then s_i is set to $+1$ if the corresponding eigenvalue is positive and -1 otherwise. The strategy for two-subgroup division can be applied to divide the entire network into multiple groups recursively if we change the label vector s into a matrix $S \in \mathbb{R}^{n \times l}$ where l is the number of groups, it starts from 1 and keeps increasing. We record the modularity before and after a new division as Q_{last} and Q_{new}, then the modularity gain is measured by $\Delta Q = Q_{new} - Q_{last}$. We stop the recursive division until there is no positive modularity gain, i.e., $\Delta Q \leq 0$. After the top-down division, we assign an unique ID to each of the detected groups based on the path from root to leaf in the hierarchical structure.

Now we address the problem of tracing the dynamic social group changes from one snapshot to the next snapshot based on the modularity maximization criteria. As time goes by, new node could be incorporated into the network and old node could also be deleted from the network. Intuitively, adding a new node that results in the insertion of one or more intra-group edges, or deleting an old node that leads to the removal of one or more inter-group edges in the current snapshot will not weaken the group structure obtained from the previous snapshot. Similarly, removing intra-group edges or inserting inter-group edges will not strengthen the group structure from the previous snapshot. However, when two groups have less distractions, adding or deleting an edge between them may change the structures of them, leading them either to merge or split further. In this case, we need to determine to which group the new node should join to maximize the modularity gain.

Inspired by an adaptive network analysis approach introduced in [183], we determine that a new node u stays in the original group C or moves to a new group C'

by two kinds of forces: $F_{stay}^{C} = e_u^C - \frac{e_u(e_C - e_u)}{2*\sum_{x_m,x_n \in TIN} e_{mn}}$ is the force to keep u stay in C and $F_{leave}^{C'} = e_u^{C'} - \frac{e_u * e_{C'}}{2*\sum_{x_m,x_n \in TIN} e_{mn}}$ is the force that C' attract u into it. Based on these two forces, the node u can determine to stay in an old group if F_{stay}^{C} is greater than any of the $F_{leave}^{C'}$, and vice versa. The proof of Theorem 1 in Appendix A demonstrates that joining the group with the largest $F_{leave}^{C'}$ will maximize the modularity gain.

Accordingly, when a node is removed in the current snapshot, it may cause a current group broken into subgroups which may further merge into other groups. To address this problem efficiently and effectively, we utilize the clique percolation method [122]. When a node is removed, a 3-clique is placed to one of its neighbor and the clique percolates until no nodes in the original group are discovered. The subgroups of original group then choose the best groups to merge. The algorithm for detecting the dynamic social groups based on the snapshots is given as Algorithm 2.

5.2.4 Unified Social Group Detection and Tracklet Fragment Association

We introduce a unified social group detection and tracklet association scheme. Sets of short tracklets extracted from two consecutive snapshots are concatenated into longer tracklets by using adapted Hungarian algorithm [90] with contextual social groups. We forward scan the tracklets until the number of non-overlapping tracklet pairs reaches the maximum number of detection responses in the frames. The starting and ending frames of the snapshot are set to the starting frame of the first tracklet and the ending frame of the last tracklet, respectively. We then restore a static version of the ETIN by including all the nodes that have a temporal overlap to the sliding window. We use the approach proposed in Section 5.2.3 to detect the social groups of tracklets and obtain the group ID for each tracklet in the time window. Finally, we integrate the

Algorithm 7: Dynamic Social Group Detection

```
Input: Current snapshot $S_{curr}$, detected social groups from previous snapshot
       $C_{pre}$: $C_1, C_2, ...C_n$, new node set $N(u)$, removal node set $R(v)$
Output: New hierarchical group structure for the current snapshot
foreach u in N(u) do
    if u has no adjacent edge then
        Create a new group with $u$ as the single member;
        Leave other groups and overall $Q$ intact;
    else if u connects existing groups then
        foreach neighbor group $C'$ do
            Calculate $F^{C'}_{leave} = e^{C'}_{u} - \frac{e_u * e_{C'}}{2 * \sum_{x_m, x_n \in TIN} e_{mn}}$;
            Find the maximum $F^{C'}_{leave}$;
        if $F^{C'}_{leave} > F^{C}_{stay}$ then
            Move $u$ to $C'$;
        else
            Leave all the groups intact;
foreach v in R(v) do
    if v.degree > $\theta$ then
        Place a 3-clique to one of $v$'s neighbor group;
        Let the clique percolate until no nodes in $C$ are discovered;
        Let the rest nodes of $C$ merge into other groups based on $Q'$;
    else
        Leave all the groups intact;
return Dynamically updated social groups;
```

group information along with the commonly used appearance and motion models into the affinity matrix M and formulate the linear assignment problem as:

$$\underset{\phi \in \Phi}{argmin} \sum_{i,j} \phi_{ij} M_{ij}, \qquad where \tag{5.7}$$

$$M_{ij} = \gamma_1[\psi_{ij} \cdot e_{ij}] + \gamma_2 f_{appr}(x_i, x_j) + \gamma_3 f_{motion}(x_i, x_j) \tag{5.8}$$

where $\psi_{ij} = lcg(x_i, x_j)/L$, L is the total number of levels of the hierarchical grouping and $lcg(\cdot, \cdot)$ is the function for computing the *lowest common group* of two tracklets in the hierarchy. Φ is the correspondence matrix with an element $\phi_{ij} = 1$ if tracklets x_i and x_j are linked and 0 otherwise. $f_{appr}(\cdot, \cdot)$ and $f_{motion}(\cdot, \cdot)$ denote the appearance and motion models, respectively. γ_is are the weighting parameters to determine the importance of each model. After the association process, the newly generated set of longer tracklets is used as input for the next round of social group detection and fragment association until all the tracks for the pedestrians are complete.

5.3 Experiments

We validate our proposed method for understanding the dynamic social grouping behavior of pedestrians on a collection of videos from real-world scenes (shopping mall, University campus, building patio) with different densities of crowds (low and medium), viewports, and sizes of the target in the frames. Sample video frames of each sequence are shown in Figure 5.5, 5.6, 5.7. Each video was recorded using elevated cameras. The videos were converted to sequences of JPEG files using the open source

Table 5.1: Percentage Distribution of the Group Sizes

	size of 1	2	3	4 or more
CAVIAR [1]	21%	57%	8%	14%
PETS2009 [2]	13%	42%	21%	24%
UNIV (newly introduced)	10%	69%	5%	16%

software “Video to Picture Converter” to produce non-interlaced 24-bit color images at a frame rate of 30 frames per second. We apply deformable part-based detectors on all the frames [56]. The detection responses from different frames are connected to form the initial short tracklets. We show and discuss how the detection responses from pedestrian detector will impact the performance of the unified social group detection and tracklet association framework in Section IV C.

5.3.1 Data Collection

The grouping information in the current video datasets is usually unavailable which requires us to manually determine the ground-truth pedestrian groups. We have manually labeled two publicly available datasets that are originally used for multi-people tracking research purpose: CAVIAR dataset [1] (low crowd density), PETS2009 [2] (medium/high crowd density). The ground-truth labeling process is conducted by asking three human judges to identify groups by assigning individuals with group IDs. The judges can rewind and play a video as many times as needed. The final consensus of the ground-truth groups are acquired by using majority voting among the judges. We also introduced a new dataset named “UNIV” which is collected from a University building patio from an elevated camera. The ground-truth for this data are also established by combining decisions made by multiple human judges. There are disagreements among the judges on some of the groups which indicates baseline ambiguity exists in the video sequences. The average disagreement rate on the number of group over the three datasets is about 5% percent of the total number of groups.

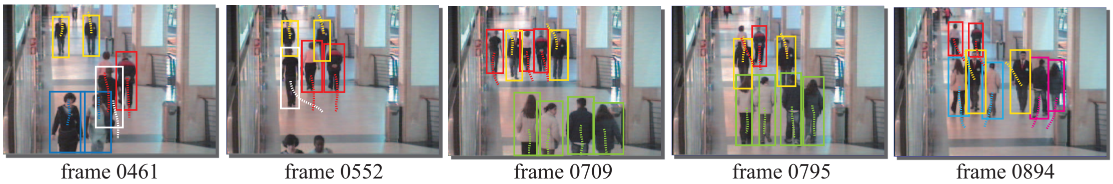
Feedback from the judges indicates that the difficulty of group identification arises when the crowd density increases. This makes PETS2009 the most difficult dataset to label. Also it is easier to identify groups from sequences with a camera viewport direction that is parallel with the walking directions of the pedestrians than the videos from camera with perpendicular views to the direction of the walking people.

CAVIAR dataset captures people walking in an indoor shopping mall environment by an elevated camera. In the dataset, people either walk from near field to far field or vice versa, and a lot of social grouping behavior can be observed during their walking. The merging and splitting of groups also happen frequently over time. There are also partial occlusions between the members of groups. PETS2009 dataset contains video sequence recorded in an outdoor scene from an University campus with a high density of people in each frame (on average 25 people are visible in each frame). Identifying individuals within a group is more challenging due to the frequent occlusions and the abrupt motion changes (direction, velocity, etc.). UNIV dataset is collected at a large camera angle under bright light conditions. The crowd density is larger than CAVIAR but smaller than PETS2009 dataset. However, more grouping behaviors and other social interactions are involved in this dataset.

The percentage distribution of group sizes from the ground-truth labeling for the video sequences are summarized in Table 5.1.

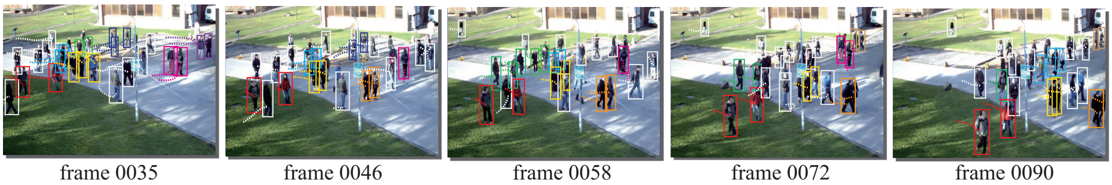
5.3.2 Quantitative Evaluation

We set ωs in Eq. (5.2) and γs in Eq. (5.8) to 1/3. We set the two thresholds $\epsilon = 0.05$ and $\epsilon' = 0.3$ in Algorithm 1. We compare the performance of our proposed



frame 0461 frame 0552 frame 0709 frame 0795 frame 0894

Figure 5.5: Group detection results from CAVIAR dataset. The pedestrians that are walking in the same group are marked in the same color. The splitting and merging behaviors are shown in the last four frames.



frame 0035 frame 0046 frame 0058 frame 0072 frame 0090

Figure 5.6: Group detection results from PETS2009 dataset. The scene is more crowded and complex with a lot of occlusions happen among pedestrians. The dynamic changes of social groups are captured by the color changes of their bounding boxes. The pedestrians with white bounding boxes are walking alone.

group detection with the following baseline approaches:

- **Baseline-I [24]**: A hierarchical agglomerative clustering based group analysis approach that starts with assigning each individual into a separate group and gradually merges the small groups into larger ones. The spatio-temporal dissimilarity between tracklets of individuals is used as a distance measure. However, it does not explicitly address the dynamic changes of social groups.

- **Baseline-II [64]**: It is another bottom-up group detection approach that is built upon algorithms for pedestrian detection and multi-people tracking. The interactions between individuals are measured by pairwise proximity and velocity without using a network representing the interactions and modularity gain as the group measure.

We compare our ETIN approach with two baseline approaches using the evaluation metric as the Percentage of correctly detected Social Groups (**PSG**) of different sizes. We measure the influence of simultaneous groups by using Percentage of cor-

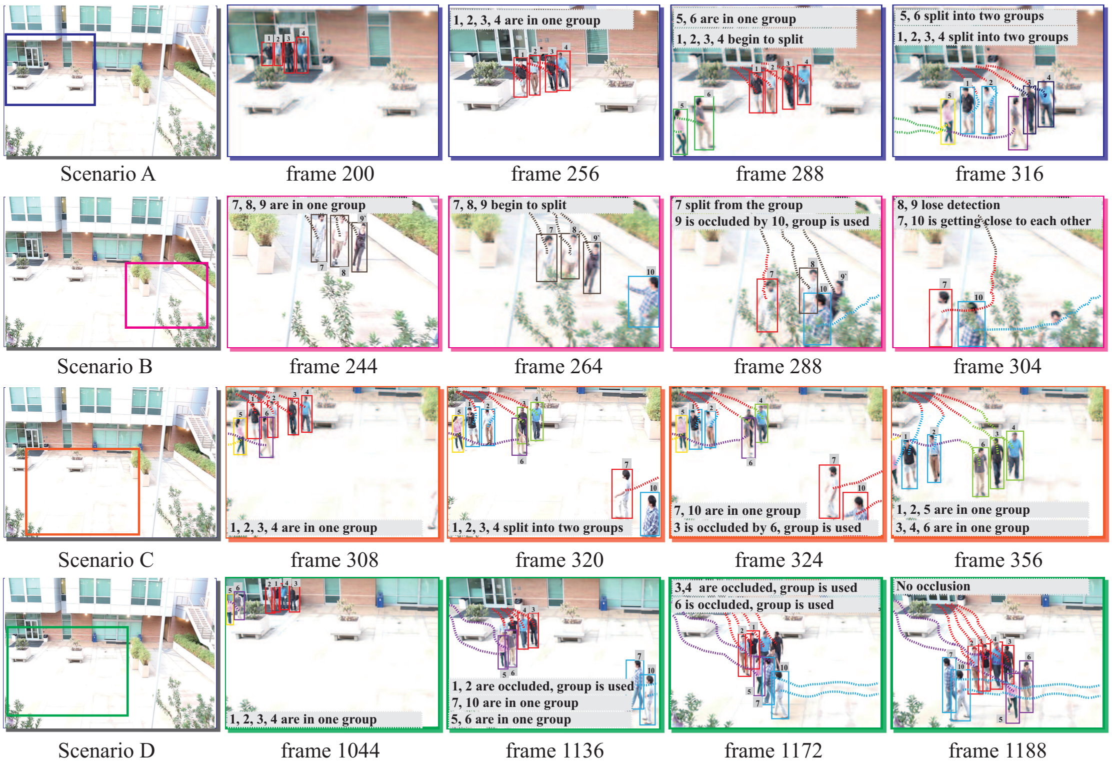
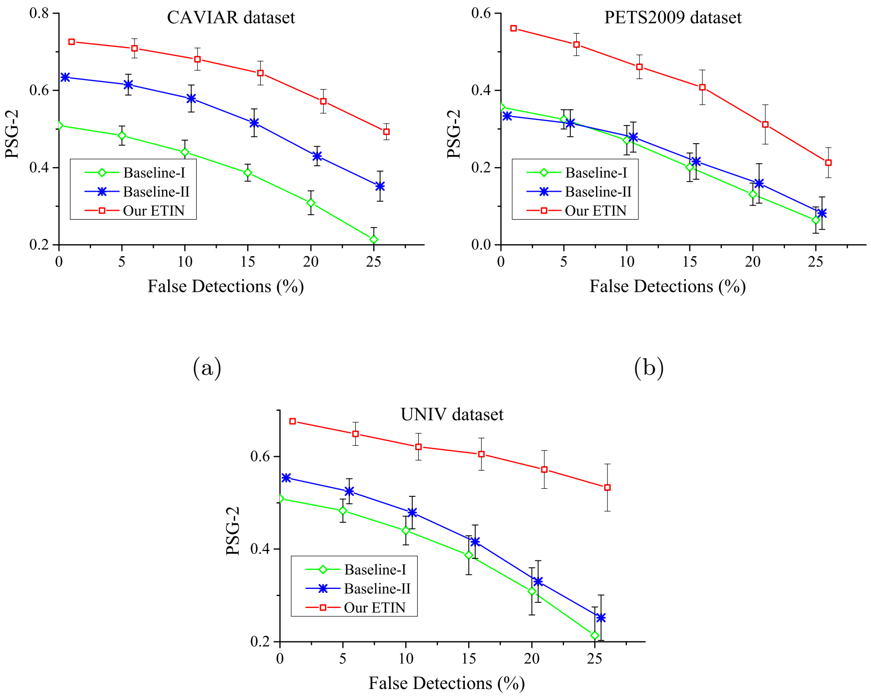


Figure 5.7: Group detection results from UNIV dataset. Group splitting and merging behaviors are shown in scenario A, B and C. Scenario D demonstrates the effect of social groups in tracklet association when partial occlusions among group members happen. The social groups are marked in different colors of bounding boxes.



(a)

(b)

(c)

Figure 5.8: The PSG measure is compared across the three approaches, as the percentage of false detections varies on (a) CAVIAR, (b) PETS2009 and (c) UNIV.

rectly detected social groups of Any size as a function of the number of Simultaneous groups (**PAS**). We also compare the performance of different approaches in tracking the dynamic social group changes (splitting and merging) by using Percentage of correctly detected Dynamic group Changes (**PDC**), which is defined as the number of correctly detected group changes by our unified detection and association approach, divided by the total number of ground-truth changes marked manually in the video frames.

- **Results on CAVIAR dataset**. We automatically detected pedestrians and generated the tracklets, and carried out the ETIN construction and modularity-based hierarchical group detection to understand the social grouping behaviors. Sample results are shown in Figure 5.5. The statistical results are summarized in Table 5.2. From the table we can observe that, all the approaches are able to identify the pedestrians walking

Table 5.2: Quantitative Evaluation on CAVIAR Dataset

Metric	Baseline-I [24]	Baseline-II [64]	our ETIN
PSG-1	67.4%	79.2%	83.5%
PSG-2	52.2%	65.7%	75.4%
PSG-3	48.5%	57.1%	69.4%
PSG-4/more	39.3%	47.6%	67.8%
PAS-1	78.5%	82.1%	86.9%
PAS-2	54.9%	62.6%	69.4%
PAS-3/more	47.3%	51.7%	61.3%
PDC	34.5%	31.2%	79.5%

alone with a high percentage of correctness. However, when the group size increases, the PSG scores from Baseline-I and Baseline-II degrades more than for our approach, which implies that our approach is more robust in detecting social groups in larger sizes. Further, when the group size is larger than 2, our PSG score is relatively stable which demonstrates the power of our network representation of tracklet interactions is stronger as compared to the pairwise social interaction representation used in other approaches. Baseline-I achieves relatively the same low score of PDC as Baseline-II which indicates that they do not actively address the dynamic group changes. This suggests that our unified framework for social group detection and tracklet association that utilizes temporal snapshots at different time intervals yields better performance in tracking the dynamic changes of social groups. Our approach achieves the best performance when more than one group appear simultaneously measured by the PAS scores. When more than two groups appears at the same time, our approach can still maintain a relatively large score (61.3%) which demonstrates that our approach can effectively handle the influences across groups.

Table 5.3: Quantitative Evaluation on PETS2009 Dataset

Metric	Baseline-I [24]	Baseline-II [64]	our ETIN
PSG-1	49.3%	53.2%	74.5%
PSG-2	39.5%	35.1%	67.3%
PSG-3	29.7%	31.7%	59.3%
PSG-4/more	29.1%	27.5%	51.6%
PAS-1	53.6%	61.7%	78.7%
PAS-2	41.9%	50.4%	63.2%
PAS-3/more	27.2%	31.3%	51.9%
PDC	26.1%	25.5%	58.4%

• **Results on PETS2009 dataset**. Similar experiments were conducted on the shorter but more challenging PETS2009 dataset. The scores of the evaluation metrics are summarized in Table 5.3. Our approach gives better results (though reduced PSGs and PDC scores) for group detection and dynamic behavior tracking performance as compared to the other two approaches. Some sample detected groups and pedestrian walking behaviors are shown in Figure 5.6. Even for this harder problem, our approach still demonstrates a substantial agreement (more than 50% of correctness) with the ground-truth not only on the different group sizes (1, 2, 3, 4 and more than 4), but also on the dynamic changes of the memberships of the groups.

A further investigation on the results shows that the PSG performance of our approach is not as stable as on the CAVIAR dataset. It degrades gradually as the number of members in the groups increase. A potential reason is that the crowd is in medium/high density and the group members tend to walk in a random pattern to avoid collisions to other pedestrians which results in a weakened social interactions among group members. For a moderate crowd of 25 people per frame, our PDC score is above

0.5, which still indicates a reasonable to good performance of our approach in tracking the dynamic group changes. The PAS scores have decreased to 51.9% compared to the CAVIAR dataset when more than two groups appear simultaneously. This implies that incorrect group information exerts on single person will have negative influence on the tracking performance when the density of the pedestrians is large and occulsion becomes a challenging problem.

• **Results on UNIV dataset**. To further evaluate the effectiveness of our approach in understanding the dynamic social groups, we applied the approach on the UNIV dataset where more social grouping behaviors are involved in a natural setting. The inter-group interactions are easier to be distinguished from the intra-group interactions in the first few frames because the groups are coming from different corners in the scene and the walking direction of each group is different. However, it becomes more challenging when the groups begins to merge and re-split in the middle frames of the video. A quantitative comparison is shown in Table 5.4.

From Table 5.4 we can observe that although the PSG scores drop to some degree compared to the scores from the CAVIAR dataset, the performance of our approach still exceeds the Baselines which demonstrates that the dynamic group analysis model and the unified group detection and tracklet association framework work effectively on this dataset where group information plays a positive role in concatenating tracklets of group members while intense occlusion happens. Overall, our proposed approach achieves the best results over the other approaches in PSG scores in all the group sizes. However, as compared to the CAVIAR dataset, the PDC scores from all the approaches have decreased to some extent as UNIV has much more dynamic social interactions that are interlaced with a large number of occlusions.

There are considerable drops in the PDC scores for the two Baselines compared

Table 5.4: Quantitative Evaluation on UNIV Dataset

Metric	Baseline-I [24]	Baseline-II [64]	our ETIN
PSG-1	60.3%	66.7%	78.6%
PSG-2	57.3%	60.4%	73.1%
PSG-3	51.2%	55.8%	70.3%
PSG-4/more	49.6%	51.4%	69.5%
PAS-1	54.8%	62.8%	74.3%
PAS-2	44.5%	49.6%	61.7%
PAS-3/more	38.6%	41.1%	48.9%
PDC	14.3%	11.9%	54.2%

to our method, particularly for the Baseline-II where the score drops from 31.2% in the CAVIAR dataset to 11.9%. The primary reason is that the other two methods do not handle group changes explicitly by investigating the group member interactions over time. The PAS score shows that our approach can still achieve a relatively good performance when more group dynamics (appear, disappear, merge, split) are involved.

5.3.3 Impact from False Detection

It is to be noted that the underlying detection errors could propagate to the group detection process in all the approaches that are based on pedestrian detection. To show that to what extent these approaches rely on accurate detection responses, we artificially introduce three types of false detections into the correct detection responses. They are: *misdetections* which represent the type of missing data, *false responses* and *inaccurate detections* that represent outliers and noises separately. The first type of false detections is added by randomly erasing correct detections and the rest two types are added by setting detections at random locations that do not cover correct detections. All the three types of false detections are added together at the percentages $[0, 5\%, 10\%, 15\%, 20\%, 25\%]$ of the total number of detections in the three datasets. The group detection performance measured by PSG-2 as a function of false detection percentages is shown in Figure 5.8.

The results from Figure 5.8 show that the robustness of our approach given unreliable detection responses. As expected, our approach maintains the best performance when the false detection percentage increases. This indicates that social groups are important contextual cues when the short tracklets are linked to form longer ones; if a group member is occluded by other pedestrians in the scene, the other group members that have close tracklet interactions can contribute to the estimation of the tracks of the occluded group member. The performance of the other two approaches that do not consider using group information in forming the trajectories drops as more false detections are obtained.

5.3.4 Application: Pedestrian Tracking

The focus of this paper is our novel approach in understanding dynamic social grouping behaviors by clustering trajectories using a social network analysis based method. However, tracking individuals by generating reliable tracks is itself a non-trivial task because of the complexity of the environment. Therefore, for completeness, we utilize our social grouping analysis framework in this section to address the individual tracking problem, which is capable of producing reasonable results that can be compared with other state-of-the-art tracking methods. Tracking individuals in the crowd is formulated as a multi-target tracking problem. We use our modified Hungarian algorithm that is integrated with individual group information to perform multi-target data association between current trajectory hypothese and the trajectories in the following

Table 5.5: Quantitative Tracking Performance on FM Dataset

Metric	DEEPER-JIGT [8]	VAR3 [8]	our ETIN
MOTP	0.80	2.80	0.31
MOTA	67.58%	2.73%	69.42%

frames (see Section III-D for more detail). Our modified Hungarian algorithm finds an optimal bipartite marching between tracklets not only based on the physical similarity but also based on the group similarity.

We evaluate our approach using the following dataset:

• **Friends Meet (FM)** [8]: contains groups of pedestrians that appear, disappear and evolve (split and merge) over time. The dataset is composed of 53 sequences for a total of 16286 frames. We use a subset of 25 sequences that contains sequences in real-life outdoor scenes. The range of the individuals in a single frame is between 3 and 11.

We use the following metrics to evaluate the performance:

• **MOTP** (Multi-Object Tracking Precision) [10]: which we define as the total error for associated tracklet-hypothesis pairs across all the time sliding windows, averaged by the total number of associations made. The value is the lower the better.

• **MOTA** (Multi-Object Tracking Accuracy) [10]: which equals one minus the mismatch rate in the data association process. It is similar to metrics widely used in other domains such as the word error rate (WER) used in speech recognition. The value is the larger the better.

We compare with the following approaches as baselines:

- **DEEPER-JIGT** (DEcentralizEd Particle filtER for Joint Individual-Group Tracking) [8]: a joint individual-group tracking framework based on decentralized particle filtering which factorizes the joint individual-group state space in two conditionally dependent subspaces. The approach is specialized in real-time tracking scenario.

- **VAR3** [8]: a variant of DEEPER-JIGT which separates individual from group tracking in two different particle filters thus blocks the contribution of the group clustring.

The results on FM dataset are summarized in Table 5.5. Our approach reaches the best performances in terms of the MOTP and MOTA evaluation. Moreover, the group information has been demonstrated as a crucial source to boosting the individual tracking. By pruning away the group information (VAR3), the performances decrease dramatically compared to other two approaches (DEEPER-JIGT and our ETIN) which build the connection between groups and individuals. In our unified social group detection and tracklet fragment association framework, the individual tracklets consider the influence from the groups in the data association process which shows the effectiveness of injecting group-driven dynamics.

5.4 Conclusions

We proposed a principled method for understanding the social grouping behavior of pedestrians as well as a unified framework for tracking the dynamic social group changes and tracklet association based on the temporal snapshots of the introduced evolving tracklet interaction network (ETIN). Our novel model addressed the social group understanding problem in video sequences from a social network perspective. The novelties included representing tracklets of pedestrians and their interactions in a network which is evolving over time and carrying out modularity to divide the tracklets into hierarchical subgroups. The dynamic changes of social groups are detected using the restored static temporal snapshots of the original network based on the time overlaps.

5.5 Proof of the Maximal Modularity Gain

***Theorem** 1: Suppose a new node u with degree d is added into the group that gives the maximum $F^{C'}_{leave}$, then adding u to C' gives the maximal modularity gain.*

Proof: Let C'' be another group of G and $C'' \neq C'$. We would like to prove that joining u into C'' will give less modularity gain than joining C'. Let $f_{C'}$ denotes the total degree of nodes in C', and let M denotes half of the summation of the total edge weights in G. The overall modularity Q when u joins C' is

$$Q = \frac{e_{C'} + e^u_{C'}}{M+d} - \frac{(f_{C'} + e^u_{C'} + d)^2}{4(M+d)^2} + \frac{e_{C''}}{M+d} - \frac{(f_{C''} + e^u_{C''})^2}{4(M+d)^2} + A \tag{5.9}$$

where A is the summation of other modularity gains. Similarly, adding u to C'' will give

$$\begin{aligned} Q' = & \frac{e_{C'}}{M+d} - \frac{(f_{C'} + e^u_{C'})^2}{4(M+d)^2} + \frac{e_{C''} + e^u_{C''}}{M+d} \\ & - \frac{(f_{C''} + e^u_{C''} + d)^2}{4(M+d)^2} + \frac{e_{C''}}{M+d} + A \end{aligned} \tag{5.10}$$

and

$$Q - Q' = \frac{1}{M+d}(e^u_{C'} - e^u_{C''} + \frac{d(f_{C''} - f_{C'} + e^u_{C''} - e^u_{C'})}{2(M+d)}) \tag{5.11}$$

since C' is the group that gives the maximum $F^{C'}_{leave}$, we have

$$e^u_{C'} - \frac{d(f_{C'} + e^u_{C'})}{2(M+d)} > e^u_{C''} - \frac{d(f_{C''} + e^u_{C''})}{2(M+d)} \tag{5.12}$$

which means

$$e^u_{C'} - e^u_{C''} + \frac{d(f_{C''} - f_{C'} + e^u_{C''} - e^u_{C'})}{2(M+d)} > 0 \tag{5.13}$$

therefore, $Q - Q' > 0$ and the conclusion is true.

Chapter 6

Summary and Future Work

In this dissertation, we proposed several methods for automated image annotation and concept-based image retrieval by exploring semantic concept co-occurrence patterns, automated moth image identification and retrieval and tracking multi-pedestrians by social groups, respectively.

In Chapter 3, we present a novel approach to automatically generate intermediate image descriptors by exploiting concept co-occurrence patterns in the pre-labeled training set that renders it possible to depict complex scene images semantically. Our work is motivated by the fact that multiple concepts that frequently co-occur across images form patterns which could provide contextual cues for individual concept inference. We discover the co-occurrence patterns as hierarchical communities by graph modularity maximization in a network with nodes and edges representing concepts and co-occurrence relationships separately. A random walk process working on the inferred concept probabilities with the discovered co-occurrence patterns is applied to acquire the refined concept signature representation. Through experiments in automatic image annotation and semantic image retrieval on several challenging datasets, we demonstrate the effectiveness of the proposed concept co-occurrence patterns as well as the concept signature representation in comparison with state-of-the-art approaches.

In Chapter 4, we describe the development of an automated moth species identification and retrieval system (SPIR) using computer vision and pattern recognition techniques. The core of the system is a probabilistic model that infers Semantically Related Visual (SRV) attributes from low-level visual features of moth images in the training set, where moth wings are segmented into information-rich patches from which the local features are extracted, and the SRV attributes are provided by human experts as ground-truth. For the large amount of unlabeled testing images in the database or added into the database later on, an automated identification process is evoked to translate the detected salient regions of low-level visual features on the moth wings into meaningful semantic SRV attributes. We further propose a novel network analysis based approach to explore and utilize the co-occurrence patterns of SRV attributes as contextual cues to improve individual attribute detection accuracy. Working with a small set of labeled training images, the approach constructs a network with nodes representing the SRV attributes and weighted edges denoting the co-occurrence correlation. A fast modularity maximization algorithm is proposed to detect the co-occurrence patterns as communities in the network. A random walk process working on the discovered co-occurrence patterns is applied to refine the individual attribute detection results. The effectiveness of the proposed approach is evaluated in automated moth identification and attribute-based image retrieval. In addition, a novel image descriptor called SRV attribute signature is introduced to record the visual and semantic properties of an image and is used to compare image similarity. Experiments are performed on an existing entomology database to illustrate the capabilities of our proposed system. We observed that the system performance is improved by the SRV attribute representation and their co-occurrence patterns.

In Chapter 5, we present a framework for characterizing hierarchical social groups based on evolving tracklet interaction network (ETIN) where the tracklets of pedestrians are represented as nodes and the their grouping behaviors are captured by the edges with associated weights. We use non-overlapping snapshots of the interaction network and develop the framework for a unified dynamic group identification and tracklet association. The approach is evaluated quantitatively and qualitatively on videos of pedestrian scenes where manually labeled ground-truth is given. The results of our approach are consistent to human-perceived dynamic social groups of the crowd. The performance analysis of our method shows that the approach is scalable and it provides situational awareness in a real-world scenarios.

Bibliography

[1] http://homepages.inf.ed.ac.uk/rbf/caviardata1/.

[2] http://www.pets2009.net/.

[3] F. Ali, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European conference on Computer vision*, pages 15–29, 2010.

[4] A. Andriyenko and K. Schindler. Decentralized particle filter for joint individual-group tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1886–1893, 2012.

[5] P. K. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[6] S. J. Bacon, S. Bacher, and A. Aebi. Gaps in border controls are related to quarantine alien insect invasions in Europe. *PLoS One*, 7(10):.doi: 10.1371/journal.pone.0047689, 2012.

[7] R. Baeza-Yates and B. Ribeiro-Neto. Modern information retrieval. *ACM Press*, pages 123–129, 1999.

[8] L. Bazzani, V. Murino, and M. Cristani. Decentralized particle filter for joint individual-group tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1886–1893, 2012.

[9] J. Berclaz, F. Fleuret, E. T. Uretken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1806–1819, 2011.

[10] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *Journal on Image and Video Processing*, 2008, 2008.

[11] B. Bhanu, R. Li, J. Heraty, and E. Murray. Automated classification of skippers based on parts representation. *American Entomologist*, pages 228–231, 2008.

[12] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10, 2008.

[13] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408, 2007.

[14] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *IEEE 15th International Conference on Computer Vision*, pages 1543–1550, 2011.

[15] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *International Conference on Computer Vision*, pages 1543 – 1550, 2011.

[16] M. Breitenstein, F. Reichlin, B. Leibe, E. K. Meier, and L. V. Gool. Robust tracking-by-detection using a detector confidence particle filter. In *IEEE 13th International Conference on Computer Vision*, pages 1515–1522, 2009.

[17] R. W. Brown. Mass phenomena. *Handbook of Social Psychology*, 2:833–876, 1954.

[18] M.L. Buffington, R.A. Burks, and L. McNeil. Advanced techniques for imaging parasitic Hymenoptera (Insecta). *American Entomologist*, 51:50–56, 2005.

[19] M.L. Buffington and M. Gates. Advanced imaging techniques ii: using a compound microscope for photographing point-mount specimens. *American Entomologist*, 54:222–224, 2008.

[20] K. Bunte, M. Biehl, M.F. Jonkman, and N. Petkov. Learning effective color features for content based image retrieval in dermatology. *Pattern Recognition*, 44:1892–1902, 2011.

[21] D. Carter. Butterflies and moths. *Eyewitness Handbooks*, 1992.

[22] A. B. Chan, Z. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[23] M. Chang, N. Krahnstoever, and W. Ge. Probabilistic group-level motion analysis and scenario recognition. In *IEEE 15th International Conference on Computer Vision*, pages 747–754, 2011.

[24] M.C Chang, N. Krahnstoever, S. Lim, and T. Yu. Group level activity recognition in crowded environments across multiple cameras. In *IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 56–63, 2010.

[25] C. W. Chen and H. Aghajan. Multiview social behavior analysis in work environments. In *ACM/IEEE International Conference on Distributed Smart Camera*, pages 1–6, 2011.

[26] X. Chen, L. An, Q. Zhen, and B. Bhanu. An online learned elementary grouping model for multi-target tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2014.

[27] M. Choi, J.J. Lim, A. Torralba, and A.S. Willsky. Exploiting hierarchical context on a large database of object categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 129 – 136, 2010.

[28] W. Choi and G. Medioni. Learning context for collective activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3273–3280, 2011.

[29] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Proceedings of the 11th European conference on Computer vision*, pages 215–230, 2012.

[30] R. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 18:370 – 383, 2007.

[31] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:603 – 619, 2002.

[32] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 886893, 2005.

[33] J. Deng, A. C. Berg, and F. Li. Hierarchical semantic indexing for large scale image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 785 – 792, 2011.

[34] C. Desai, D. Ramanan, and C.C. Fowlkes. Discriminative models for multi-class object layout. *International Journal on Computer Vision*, 95:1–12, 2011.

[35] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien. Semtag and seeker: Bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the 20th International Conference on World Wide Web*, pages 178–186, 2003.

[36] L. Ding and A. Yilmaz. Learning relations among movie characters: a social network perspective. In *Proceedings of the 11th European conference on Computer vision*, pages 410–423, 2010.

[37] L. Ding and A. Yilmaz. Inferring social relations from visual concepts. In *IEEE 15th International Conference on Computer Vision*, pages 699–706, 2011.

[38] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1271 – 1278, 2009.

[39] M. T. Do, J. M. Harp, and K. C. Norris. A test of a pattern recognition system for identification of spiders. *Bulletin of Entomological Research*, 89(3):217–224, 1999.

[40] A. Dong and B. Bhanu. Active concept learning in image databases. *IEEE Trans. on Systems, Man and Cybernetics, Part B*, 35:450–456, 2005.

[41] M. Douze, A. Ramisa, and C. Schmid. Combining attributes and fisher vectors for efficient image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 745 – 752, 2011.

[42] M. Douze, A. Ramisa, and C. Schmid. Combining attributes and fisher vectors for efficient image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 745 – 752, 2011.

[43] M. Douze, A. Ramisa, and C. Schmid. Combining attributes and fisher vectors for efficient image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 745 – 752, 2011.

[44] K. Duan, D. Parikh, and D. Crandall. Discovering localized attributes for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3474 – 3481, 2012.

[45] R. Eshel and Y. Moese. Homography based multiple camera detection and tracking of people in a dense crowd. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[46] E. Estrada and N. Hatano. A vibrational approach to node centrality and vulnerability in complex networks. *Physica A: Statistical Mechanics and its Applications*, 389:36483660, 2010.

[47] D. Fernndez V. Formoso F. Cacheda, V. Carneiro. Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Transactions on the Web*, 5, 2011.

[48] J. Fan, Y. Gao, and H. Luo. Hierarchical classification for automatic image annotation. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 111–118, 2007.

[49] J. Fan, Y. Gao, and H. Luo. Mining multilevel image semantics via hierarchical classification. *IEEE Transactions on Image Processing*, 71:407–426, 2008.

[50] J. Fan, Y. Gao, H. Luo, and R. Jain. Mining multilevel image semantics via hierarchical classification. *IEEE Transactions on Multimedia*, 10:167–187, 2008.

[51] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2352–2359, 2010.

[52] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2352 – 2359, 2010.

[53] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785, 2009.

[54] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778 – 1785, 2009.

[55] A. Fathi, J.K. Hodgins, and J.M. Rehg. Social interactions: A first-person perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233, 2012.

[56] P. Felzenszwalb, D. McAllester, and D. Ramaman. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[57] L. Feng and B. Bhanu. Utilizing co-occurrence patterns for semantic concept detection in images. In *IEEE 21st International Conference on Pattern Recognition*, pages 2918–2921, 2012.

[58] A. Foncubierta-Rodrguez, A. G. S. Herrera, and H. Müller. Medical image retrieval using bag of meaningful visual words: unsupervised visual vocabulary pruning with plsa. In *Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare*, pages 75–82, 2013.

[59] S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75174, 2010.

[60] T. M. Francoy, D. Wittmann, M. Drauschke, S. Müller, V. Steinhage, M. A. F. Bezerra-Laure, D. D. Jong, and L. S. Goncalves. Identification of africanized honey bees through wing morphometrics: two fast and efficient procedure. *Apidologie*, 39(5):488–494, 2008.

[61] G. Fu, F.Y. Shih, and H. Wang. A kernel-based parametric method for conditional density estimation. *Pattern Recognition,*, 44(2):284–294, 2011.

[62] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[63] T. Ganchev, I. Potamitis, and N. Fakotakis. Acoustic monitoring of singing insects. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, 2007.

[64] W. Ge, R.T. Collins, and B. Ruback. Automatically detecting the small group structure of a crowd. In *IEEE Workshop on Applications of Computer Vision*, pages 1–8, 2009.

[65] W. Ge, R.T. Collins, and R.B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:1003–1016, 2012.

[66] J. C. Van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271 – 1283, 2010.

[67] G. Gennari and G. Hager. Probabilistic data association methods in visual tracking of groups. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 876–881, 2004.

[68] N. Ghosh and B. Bhanu. Evolving bayesian graph for 3d vehicle model building from video. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–16, 2013.

[69] M. Girvan and M. E. Newman. Community structure in social and biological networks. In *Proceedings of the National Academy of Science*, page 78217826, 2002.

[70] D. Goldberg, D. Nichols, B.M. Oki, and D.B. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35:61–70, 1992.

[71] C. C. Gotlieb and H. E. Kreyszig. Texture descriptors based on co-occurrence matrices. *Computer Vision, Graphics, and Image Processing,*, 51(1):70–86, 1990.

[72] Stephen Gould, Tianshi Gao, and Daphne Koller. Region-based segmentation and object detection. In *Advances in Neural Information Processing Systems*, pages 655–663, 2009.

[73] A. Hanjalic, R. Lienhart, W. Y. Ma, and J. R. Smith. The holy grail of multimedia information retrieval: So close or yet so far away? *Proceedings of the IEEE*, 96(4):541–547, 2008.

[74] R. M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE,*, 67:786–804, 1979.

[75] I. Haritaoglu and M. Flickner. Detection and tracking of shopping groups in stores. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 431–438, 2001.

[76] W. H. Hsu, Lyndon S. Kennedy, and S-F. Chang. Video search reranking through random walk over document-level context graph. In *Proceedings of the 15th International Conference on Multimedia*, pages 971–980, 2007.

[77] Y. Hu, M. Li, P. Zhang, Y. Fan, and Z. Di. Community detection by signaling on complex networks. *Physical Review E*, 78, 2008.

[78] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *Proceedings of the 9th European conference on Computer vision*, pages 788–801, 2008.

[79] S. Hwang and K. Grauman. Reading between the lines: Object localization using implicit cues from image tags. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:1145 – 1158, 2012.

[80] D. H. Janzen and W. Hallwachs. Dynamic database for an inventory of the macrocaterpillar fauna, and its food plants and parasitoids, of area de conservacion guanacaste (acg), northwestern costa rica (nn-srnp-nnnnn voucher codes), 2009. Available from http://janzen.sas.upenn.edu.

[81] U. Jean, D. Arne, F. Simon, and G. Malcolm. A stile project case study: The evaluation of a computer-based visual key for fossil identification. *Association for Learning Technology Journal*, 4(2):40–47, 1996.

[82] Y. G. Jiang and A.G. Hauptmann J. Yang, C. W. Ngo. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia*, 12(1):42–53, 2010.

[83] Y. G. Jiang, C. W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 494–501, 2007.

[84] A. Joly, H. Goau, H. Glotin, C. Spampinato, P. Bonnet, W.P. Vellinga, R. Planque, A. Rauber, R. Fisher, and H. Müller. Lifeclef 2014: multimedia life species identification challenges. In *Proceedings of the LifeCLEF 2014*, pages 229–249, 2014.

[85] S-H. Kang, W. Jeon, and S-H. Lee. Butterfly species identification by branch length similarity entropy. *Journal of Asia-Pacific Entomology*, 15(3):437–441, 2012.

[86] Y. Kaya and L. Kayci. Application of artificial neural network for automatic detection of butterfly species using color and texture features. *The visual computer*, 30(1):71–79, 2014.

[87] P.H. Kerr, E.M. Fisher, and M.L. Buffington. Dome lighting for insect imaging under a microscope. *American Entomologist*, 54:198–200, 2008.

[88] P. Kilamba, E. Ribnick, A. Joshi, O. Masoud, and N. Papanikolopoulos. Estimating pedestrian counts in groups. *Computer Vision and Image Understanding*, 110:43–59, 2008.

[89] K.Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60:6386, 2004.

[90] H.W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2, 1955.

[91] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE 13th International Conference on Computer Vision*, pages 365–372, 2009.

[92] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *International Conference on Computer Vision*, pages 365 – 372, 2009.

[93] S. Kumschick, S. Bacher, W. Dawson, and J. Heikkil. A conceptual framework for prioritization of invasive alien species for management according to their impact. *NeoBiota*, 15(10):69–100, 2012.

[94] F. L, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:594611, 2006.

[95] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 951 – 958, 2009.

[96] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 951 – 958, 2009.

[97] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.

[98] Z. Li, J. Liu, X. Zhu, T. Liu, and H. Lu. Image annotation using multi-correlation probabilistic matrix factorization. In *Proceedings of the international conference on Multimedia*, pages 1187–1190, 2010.

[99] H. Ling and K. Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 29(5):840–853, 2007.

[100] D. Liu, X. Hua, L. Yang, M. Wang, and H. Zhang. Tag ranking. In *Proceedings of the 18th international conference on World wide web*, pages 351–320, 2009.

[101] Y. Liu, R. Jin, and L. Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *Proceedings of the 21st national conference on Artificial intelligence*, pages 421–426, 2006.

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

[103] X. Luo, M. Zhou, Y. Xia, and Q. Zhu. An efficient non-negative matrix-factorization-based approach to collaborative-filtering for recommender systems. *IEEE Transactions on Industrial Informatics*, 10:1273 – 1284, 2014.

[104] H. Ma, J. Zhu, M. R-T. Lyu, and I. King. Bridging the semantic gap between image contents and tags. *IEEE Transactions on Multimedia*, 12:462–473, 2010.

[105] T. Malisiewicz, A. Gupta, and A.A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *IEEE 15th International Conference on Computer Vision*, pages 89 – 96, 2011.

[106] M. Mayo and A. T. Watson. Automatic species identification of live moths. *Knowledge-Based Systems*, 20(4):195–202, 2007.

[107] C. McPhail. Withs across the life course of temporary sport gatherings. In *Univ. of Illinois*, 2003.

[108] C. McPhail and R. Wohlstein. Using film to analyze pedestrian behavior. *Sociological Meth. and Research*, 10:347–375, 1982.

[109] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942, 2008.

[110] T. D. Meulemeester, P. Gerbaux, M. Boulvin, A. Coppe, and P. Rasmont. A simplified protocol for bumble bee species identification by cephalic secretion analysis. *International Journal for the Study of Social Arthropods*, 58(5):227236, 2011.

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

[112] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[113] M. Mitrovic and B. Tadic. Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities. *Physical Review E*, 80, 2009.

[114] M. Moussaid, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz. The walking behavior of pedestrian social groups and its impact on crowd dynamics. *PLos ONE*, 5, 2010.

[115] H. Müller, T. Clough, P. Deselaers, and B. Caputo. *ImageCLEF Experimental evaluation in visual information retrieval.* Springer, 2010.

[116] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69, 2004.

[117] M. E. Newman and M. Girvan. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74, 2006.

[118] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, 2004.

[119] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E,*, 69(2):84–99, 2004.

[120] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

[121] M.A. ONeill, I.D. Gauld nd K.J. Gaston, and P.J.D. Weeks. Daisy: an automated invertebrate identification system using holistic vision techniques. In *Inaugural Meeting of the BioNET-International Group for Computer-aided Taxonomy*, 2000.

[122] G. Palla, P. Pollner, A. Barabasi, and T. Vicsek. Social group dynamics in networks. *Adaptive Networks*, pages 11–38, 2009.

[123] G. Pallal, I. Dernyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, 2005.

[124] X. Pan, C.S. Han, K. Dauber, and K.H. Law. A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations. In *AI Society*, pages 113–132, 2007.

[125] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1681 – 1688, 2011.

[126] D. Parikh and K. Grauman. Relative attributes. In *IEEE 15th International Conference on Computer Vision*, pages 801–808, 2011.

[127] D. Parikh and K. Grauman. Relative attributes. In *International Conference on Computer Vision*, pages 801 – 808, 2011.

[128] S. Pellegrini, A. Ess, and L.V. Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *Proceedings of the 9th European conference on Computer vision*, pages 452–465, 2010.

[129] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. Youll never walk alone: Modeling social behavior for multi-target tracking. In *IEEE 13th International Conference on Computer Vision*, pages 261–268, 2009.

[130] H. M. Pereira, S. Ferrier, M. Walters, G. N. Geller, R. H. G. Jongman, R. J. Scholes, M. W. Bruford, N. Brummitt, S. H. M. Butchart, A. C. Cardoso, and et al. Essential biodiversity variables. *Science*, 339(1):277278, 2013.

[131] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[132] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirie. Large-scale image retrieval with compressed fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3384 – 3391, 2010.

[133] J. Philbin, O. Chum, M. Isard, and J. Sivic. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[134] V. S. N. Prasad and B. Yegnanarayana. Finding axes of symmetry from potential fields. *IEEE Transactions on Image Processing*, 13(12):1559–1566, 2004.

[135] Z. Qin and C. Shelton. Improving multi-target tracking via social grouping. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1972–1978, 2012.

[136] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[137] V. Ramanathan, B. Yao, and F. Li. Social role discovery in human events. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2475–2482, 2013.

[138] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth movers distance as a metric for image retrieval. *International Journal of Computer Vision,*, 40(2):99–121, 2000.

[139] Y. Rubnera, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40, 2000.

[140] B. C. Russell and et al. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:1–3, 2008.

[141] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, 2008.

[142] M. Ryoo and J. Aggarwal. Recognition of high-level group activities based on activities of individual members. In *IEEE Workshop on Motion and Video Computing*, pages 1–8, 2008.

[143] K. E. A. Van De Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582 – 1596, 2010.

[144] S. Schroder, W. Drescher, V. Steinhage, and B. Kastenholz. An automated method for the identification of bee species. In *Proceedings of the International Symposium on Conserving European Bees*, pages 6–7, 1995.

[145] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 – 8, 2007.

[146] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proceedings of the 7th European conference on Computer vision*, pages 1–15, 2006.

[147] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 801 – 808, 2011.

[148] B. Siddiquie, R.S Feris, and L.S Davis. Image ranking and retrieval based on multi-attribute queries. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 801 – 808, 2011.

[149] N. Singhai and S.K. Shandilya. A survey on: content based image retrieval systems. *International Journal of Computer Applications*, 4:22–26, 2010.

[150] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. In *International Conference on Computer Vision*, pages 1543 – 1550, 2005.

[151] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):13491380, 2000.

[152] B. Solmaz, B. E. Moore, and M. Shah. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:20642070, 2012.

[153] B. Song, T.Y. Jeng, E. Staudt, and A.K. Roy-Chowdhury. A stochastic graph evolution framework for robust multi-target tracking. In *Proceedings of the 11th European conference on Computer vision*, pages 605–619, 2010.

[154] P. R. Steele and J. C. Pires. Biodiversity assessment: State-of-the-art techniques in phylogenomics and species identification. *American Journal of Botany*, 98(3):415–425, 2011.

[155] Y. Sun and B. Bhanu. Reflection symmetry-integrated image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1827–1841, 2012.

[156] P. Tirilly, V. Claveau, and P. Gros. Language modeling for bag-of-visual words image categorization. In *International Conference on Content-based Image and Video Retrieval*, pages 249–258, 2008.

[157] A. Tofilski. Drawwing, a program for numerical description of insect wings. *Journal of Insect Science*, 4:17–22, 2004.

[158] A. Torralba. Contextual priming for object detection. *International Journal on Computer Vision*, 53:169–191, 2003.

[159] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:19581970, 2008.

[160] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *Advances in Neural Information Processing Systems*, pages 1401–1408, 2005.

[161] L. Torresani, M. Summer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *Proceedings of the 7th European conference on Computer vision*, pages 776–789, 2010.

[162] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu. Homography based multiple camera detection and tracking of people in a dense crowd. In *European Conference on Computer Vision*, pages 691–704, 2008.

[163] S. Uchihashi and T. Kanade. Content-free image retrieval by combinations of keywords and user feedbacks. *Image and Video Retrieval*, 3568:650–659, 2005.

[164] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha. Real-time bag of words, approximately. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval*, pages 494–501, 2009.

[165] O. Vicente, G. Kulkarni, and T.L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, 2011.

[166] V. Viitaniemi and J. Laaksonen. Spatial extensions to bag of visual words. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 197–206, 2009.

[167] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[168] J. Wang, L. Ji, A. Liang, and D. Yuan. The identification of butterfly families using content-based image retrieval. *Biosystems Engineering*, 111:24–32, 2011.

[169] J. Wang, C. Lin, L. Ji, and A. Liang. A new automatic identification system of insect images at the order level. *Knowledge-Based Systems*, 33:102–110, 2012.

[170] S. Wang, H. Lu, F. Yang, and M. H. Yang. Robust tracking-by-detection using a detector confidence particle filter. In *IEEE 15th International Conference on Computer Vision*, pages 1323–1330, 2011.

[171] X. Wang, K. Liu, and X. Tang. Query-specific visual semantic spaces for web image re-ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 857 – 864, 2011.

[172] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:539–555, 2009.

[173] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *Proceedings of the 11th European conference on Computer vision*, pages 155–168, 2010.

[174] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *European Conference on Computer Vision*, pages 155–168, 2010.

[175] Y. D. Wang, J. K. Wu, A. A. Kassim, and W. M. Huang. Tracking a variable number of human groups in video using probability hypothesis density. In *International Conference on Pattern Recognition*, pages 1127–1130, 2006.

[176] C. Wen, D. E. Guyer, and W. Li. Local feature-based identification and classification for orchard insects. *Biosystems Engineering*, 104(3):299–307, 2009.

[177] P. Wilkins, P. Ferguson, and A. F. Smeaton. Using score distributions for query-time fusion in multimediaretrieval. *Multimedia information retrieval*, 2006.

[178] Y. Xiang, X. Zhou, Z. Liu, T-S. Chua, and C-W Ngo. Semantic context modeling with maximal margin conditional random fields for automatic image annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3368 – 3375, 2010.

[179] K. Yamaguchi, A. Berg, L. Ortiz, and T. Berg. Who are you with and where are you going? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1345–1352, 2011.

[180] B. Yang and R. Nevatia. An online learned crf model for multi-target tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2034–2041, 2012.

[181] J. Yang, Y. G. Jiang, A. G. Hauptmann, and C. W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on multimedia information retrieval*, pages 197–206, 2007.

[182] Y. Yang and S. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the International Conference on Advances in Geographic Information Systems*, pages 270–279, 2010.

[183] Z. Ye, S. Hu, and J. Yu. Adaptive clustering algorithm for community detection in complex networks. *Physical Review E*, 78, 2008.

[184] P. Y. Yin, B. Bhanu, and K. C. Chang. Long-term cross-session relevance feedback using virtual features. *IEEE Transactions on Knowledge and Data Engineering*, 20(3):352 – 368, 2008.

[185] P. Y. Yin, B. Bhanu, K. C. Chang, and A. Dong. Long-term cross-session relevance feedback using virtual features. *IEEE Transactions on Knowledge and Data Engineering,*, 20(3):352–368, 2008.

[186] F.X. Yu, R. Ji, M.H. Tsai, G. Ye, and Shih-Fu Chang. Weak attributes for large-scale image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2949 – 2956, 2012.

[187] Q. Yu and G. Medioni. Multiple-target tracking by spatio-temporal monte carlo markov chain data association. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[188] T. Yu, S. N. Lim, K. Patwardhan, and N. Krahnstoever. Monitoring, recognizing and discovering social networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1462–1469, 2009.

[189] T. Yu, S. N. Lim, K. Patwardhan, and N. Krahnstoever. Monitoring, recognizing and discovering social networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1462–1469, 2009.

[190] J. Yue, Z. Li, L. liu, and Z. Fu. Content-based image retrieval using color and texture fused features. *Mathematical and Computer Modeling*, 54:1121–1127, 2011.

[191] B. Zhan, D.N. Monekosso, P. Remagnino, S. A. Velastin, and L. Xu. Crowd analysis: A survey. *Journal of Machine Vision and Applications*, 19:345–357, 2008.

[192] J. Zhang, M. Marsza lek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73(2):213238, 2007.

[193] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[194] S. Zhang, R. S. Wang, and X. S. Zhang. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications*, 374:483490, 2007.

[195] W. Zhang, F. Chen, W. Xu, and Y. Du. Hierarchical group process representation in multi-agent activity recognition. *Image Communication*, 23:739–739, 2008.

[196] Y. Zhang, W. Ge, M.C. Chang, and X. Liu. Group context learning for event recognition. In *IEEE Workshop on Applications of Computer Vision*, pages 249–255, 2012.

[197] T. Zhao, R. Nevatia, and B. Wu. Segmentation and tracking of multiple humans in crowded environments. *IEEE Transactions on Pattern Analysis and Machine and Intelligence*, 30:1198–1211, 2008.

[198] N. Zhou, W.K. Cheung, G. Qiu, and X. Xue. A hybrid probabilistic model for unified collaborative and content-based image tagging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1281–1294, 2011.

[199] L.Q. Zhu and Z. Zhang. Auto-classification of insect images based on color histogram and GLCM. In *7th International Conference on Fuzzy Systems and Knowledge Discovery*, pages 971–980, 2010.

[200] R. Van Zwol. Multimedia strategies for b3-sdr, based on principal component analysis. *Lecture notes in computer science*, 3977:540553, 2005.