# Frontiers in Image and Video Analysis
# NSF/FBI/DARPA Workshop Report

Rama Chellappa

# Contents

# 1 Executive summary

## 1.1 Motivation

The bombing attacks at the Boston Marathon in April 2013 and recent incidents at the Navy Yard and the shopping center in Nairobi, Kenya, presented the law enforcement community with significant challenges in terms of the volume and variety of video and still images acquired in the course of the investigation. Tens of thousands of individual media files in multiple formats were submitted from a variety of sources. These sources included broadcast television feeds, private Close-Circuit Television (CCTV) systems, mobile device photographs and videos recovered from the scene, as well as photographs and videos submitted by the public. Teams of analysts reviewed this evidence using mostly manual processes to determine the sequence of events before and after the bombing, ultimately leading to a quick resolution of the case. In the aftermath, it has become evident that the proliferation of video and image recording devices in fixed and mobile devices make it inevitable that a similar situation will occur in future events. As a result, it is incumbent upon the law enforcement community and the U.S. Government at large to further explore the use of automated approaches, available today or in the coming years, to better organize and analyze such large volumes of multimedia data.

The workshop was organized to establish a better understanding of the state of the art in algorithms being developed in academia that can support forensic analysis and identification in large volumes of images and videos (e.g., multimedia). The stakeholders (funding agencies and operational entities) must make plans for long- and near-term research and development efforts if they are to be prepared to optimally address this situation in the future. The workshop organizers and attendees were charged to identify those video and image analysis problems which are: (1) Considered solved (i.e., ready to deploy in specific operational scenarios): (2) Nearly solved (i.e., 1-to-3 years to deployment); and (3) Over-the-Horizon problems (i.e., those requiring concerted effort over the next 3-5 years and beyond).

In order to provide a structure for the expected discussions, the workshop was divided into nine sessions. These sessions were: video summarization, shot detection/scene change detection, geo-tagging, image-based biometrics, role of humans, human action recognition, re-identification, semantic description, large-scale image recognition and data collection and performance evaluation. In each session, invited panelists gave brief views on the three major themes (solved, nearly solved and over-the-horizon problems) followed by discussions among the participants present in the sessions. On the second day, the workshop organizers summarized the discussions from all the nine sessions. The agenda of the workshop is in Appendix A.

The workshop attracted about eighty participants form universities, industries, FFRDCs, government laboratories and other entities with interests in topics discussed at the workshop.

## 1.2 Findings

Based on discussions at the workshop and subsequent conversations, the attendees arrived at lists of problems that are seen as solved and problems that require near-term and long-term investments. These problems are discussed in great detail in chapters 2-10. In this section

we provide a summary of problems that need long-term investments.

### 1.2.1 Problems that Need Long Term Investments

**Video Summarization:** Video summaries that are generated in response to user-specified set of rules that can be computationally interpreted and translated to image/video operations are desired. Video summaries that are robust to poor spatial resolution of objects of interest, noise and jitter, and poor illumination conditions must be developed. Current visualization techniques are mostly limited to visualizing a single camera stream. Summaries that come from multiple viewpoints, perhaps even city-wide camera networks may be needed to generate a complete picture of what is being imaged. We need to invest in designing 2.5D summarization techniques that in addition to summarizing 2D spatial video, can bring into consideration overlapping views, and qualitative geometric models of wide area scenes. Metrics for evaluating a good summary are needed. The quality of summarization needs to be studied in-situ, in a deployed environment, with end-users providing qualitative feedback on their experience with these systems.

**Visual Analysis and Geo-localization of Large-Scale Imagery:** In the longer term algorithms and systems that would (semi-)automatically determine the level of precision achievable for a given geo-location problem and then apply the appropriate methods to get to one of the three precision regimes: visual element location, region location, pinpoint location must be developed. When directly matchable, or even regionally matchable attributes are not available in imagery, methods that derive correlations between visual appearance and location using large scale data will be required. In order to derive such correlations, mid-level and high-level (towards semantics) descriptions of images will need to be computed. This is the domain of problems where recognition in the large meets geo-location. In the domain of semantics based correlation, we will need to create human like capabilities for reasoning with and matching image attributes.

**Image-based Biometrics:** Two fundamental problems in biometrics are coping with imagery acquired in unconstrained and challenging environments, and dealing with the effects of the passage of time on a biometric. In the arena of face recognition, recognizing an individual seen in challenging viewing conditions, such as in extremely low-resolution images, or recognizing a person from an extreme viewpoint, such as a profile, when only a frontal view is present in the gallery are problems that need long-term investments. Changes in appearance over a long period of time (aging) are more challenging, and less well understood, because these variations are more difficult to model accurately. Additionally, significant changes can occur even over a very short period of time, as people gain or lose weight or change their eye ware or facial hair. These short-term changes are also not yet very well modeled by existing approaches. Further, there is a lack of representations that are invariant to the passage of time for human faces, in the same way that they have been developed for fingerprints. The problem is further confounded when one or more of these variations are observed simultaneously in the face image.

For the general biometrics problem, methods for computing the degree of permanence of a biometric trait/template must be developed. Likewise, models for characterizing the changes

over time of a biometric trait/template must be developed. Robust biometrics methods that can handle signatures acquired in unconstrained sensing environments must be developed. These include faces in surveillance videos, iris in motion and latent fingerprints in crime scenes.

Methods for designing secure biometric signatures are still at their infancy. Biometric methods that ensure system security and user privacy are needed. Most work on face recognition has attempted to identify a face using a cropped window of the image containing a single face. However, in many cases, using a larger context may assist in recognition. For example, some people may tend to appear together, so the identity of one person may provide evidence about the identity of another. Sublinear methods for searching over large databases using descriptive features such as attributes must be developed. Recent advances in technology have made it possible to build very large labeled image sets. Results using proprietary dataset of several million labeled faces are being released. However, it may become possible in the near future to build even larger image sets. New research is needed both to determine how to build these image sets efficiently and to understand how such image sets can be used most effectively.

**Human in the Loop:** There are several long-term questions for the HIL problem. Many of these originate in open questions in the area of visual analysis, which is itself a widely open problem at this point. However, in many cases, these questions are also central to other areas of HIL, such as perceptual modeling or interactivity. Humans think along many dimensions, fusing visual, auditory, memory based (priors), and tactile information, among others. Central to this process is the establishment of cross-modal representations, i.e. representations which abolish barriers across modalities, bringing information from all the modalities into a universal coordinate frame. To be truly cross-modal, a representation must support inference even when some of the modalities are hidden or unavailable. Support for this type of inference could lead to major breakthroughs for the HIL problem.

Humans can quickly transfer knowledge from one task to the next, from one modality to another, etc. This is a central requirement for the human ability for zero-shot learning, i.e. learning new concepts from a very limited number of examples. However, current HIL systems provide very little support for this type of functionality. This problem is present in all areas of HIL, from vision systems that cannot transfer information across camera views, to interfaces that cannot transfer information across user sessions, to perceptual models that cannot learn the commonalities and intricacies of different users. While there has been some progress in all these problems, both the theoretical and algorithmic foundations are still in their infancy. It is also only now that enough computation and storage are becoming available for researchers to worry about problems such as dataset bias, i.e. how well an algorithm learned under certain training conditions generalizes to others.

Humans abstract information into semantic representations involving abstract concepts. While semantic representations have long been used in multimedia and more recently in vision, there are still a number of hard open questions in this field. One obvious limitation is that most semantic representations are flat, i.e. assume that all concepts have equal semantic level. While a few efforts have attempted to use taxonomies, these are usually hard-coded, e.g. based on Wordnet. On the other hand, human taxonomies are a mix of common sense knowledge and user experience, acquired through years of interaction with the world. There

is currently very little ability to learn a taxonomy, or adapt an existing taxonomy to a user, e.g. based on the patterns of interaction with an IVA system. There has also been little work in trying to define semantic metric structures that truly mimic those used in human similarity judgments, or to design cross-modal systems that exploit these structures to transfer knowledge across information modalities. These, and most other problems in the area of semantic representations, are likely to become more central in the coming decades, as large quantities of storage and computing become available, enabling the routine design of semantic representations with thousands of concepts.

The study of biological vision systems, the development of computational models that explain the functionality of these systems, and the translation of these models into computer vision algorithms that solve multiple vision tasks, should be the subject of much heavier emphasis in computer vision than they are today. While the recent excitement about deep learning has some of these characteristics, current deep learning models are far from realistic models of biological computing. Critical information processing components, such as 1) the feedback from higher cognition necessary to implement top-down vision routines or 2) different types of normalization, are simply not accounted for. A better understanding of these components, as well as of the principles that guide biological computation, could thus prove critical to enable universal vision systems.

The current paradigms for learning from and responding to user feedback are based on the notion of risk minimization. This is an estimate of average system performance, which works well for problems such as classification and regression and is quite popular in machine learning. Most on-line learning or active learning algorithms, on which HIL systems depend, are based on these principles. However, in the HIL context, optimal performance on average is not necessarily satisfying. Learning methods that optimize worst-case instead of average performance are needed.

Machine learning research has overwhelming addressed the problems of supervised (classification, regression) and unsupervised (clustering, deep network) learning. Yet, in the HIL context, virtually all problems lie in between these two extremes, or what is usually called weakly supervised learning. This is because supervision tends to be tedious and cannot ever be acquired in full detail. Investments in the theory of weakly-supervised learning and the consequent development of learning algorithms with better performance guarantees are thus needed for the HIL area.

**Person Re-Identification:** Full real-world scenarios, low-quality images, unconstrained and uncooperative conditions are challenges that will need a longer time horizon. This will involve dealing with natural videos with high clutter in the data, and severe variations in the environmental conditions. An example could be a busy city scene, where a person needs to be recognized as he walks through several blocks with large blind areas in between. The main task - robust feature extraction - remains the key and will need to be achieved in far more challenging conditions. This may call for the development of novel features. Tools that could be useful for this purpose include image restoration (e.g., super resolution) techniques to improve the quality of the acquired images/videos. It is to be expected that a larger use of context, like the joint re-id of groups and individuals, can be helpful. Semantically meaningful attributes could play a role in providing the required robustness. The development of online learning can be a major step in the feature extraction problem. Methods (like deep

learning and sparse coding) which can automatically learn the best features, rather than using hand-engineered ones, hold promise in this respect. The use of multi-modal multi-sensory input (beyond optical, e.g., multi-spectral, infrared) should also be considered to cope with real scenarios, as well as the potential exploitation of active acquisition systems and mobile platforms for collecting more information-rich data. This could allow capturing salient parts of the human body as a way to acquire features that would enable recognizing people in such harsh scenarios. All these techniques should scale gracefully with large numbers of cameras, and cope with wider space-time horizons.

The construction of large datasets in the wild will be a necessity to validate the developed re-id technologies. However, collecting such datasets that will be meaningful and annotating them reliably will be a challenge.

**Human Activity Understanding (Detection and Recognition) in a Video:** The ultimate goal of activity and action understanding is to be able to provide explanations and descriptions of an action or event captured in a video. This kind of analysis is very important for end users, e.g., video analysts. To understand and explain a video, it is important to have a rich representation of each event, action or activity in terms of objects, actions and scenes, which can be used to describe an event in natural language. Investments to support efforts that aim to bridge the gap between the semantics required by the high level description and what can be extracted from the lower level detectors are needed. Given that great progress has been made in robust extraction of low level features it will be interesting to revive model and knowledge based approaches for understanding human activities. The long term challenge for computer vision researchers is to develop approaches for human activity and action understanding which do not require any training and which can generalize to diverse datasets and provide explanation and recounting of actions and activities in a video.

**Semantic Summarization (Attribute-based Scene Tagging):** Imagery will increasingly provide higher resolution data over broader fields-of-view to stress the generally bottom-up processing of semantic summarization processing. That is, it will become increasingly necessary to invest in top-down scene understanding to develop interpretations of aggregations of objects along with the conventional processing framework that builds interpretations from pixels to regions to parts to objects to attributes to relationships. New challenges include the direct detection of groups and crowds and characterizing their approximate counts, distribution of appearance attributes and distribution of poses.

Conventional semantic summarization techniques apply a range of feature extraction algorithms to identify scene regions, objects and scene geometry. But semantics are not only based on appearance properties, but also on function such as, what a person is doing with a tool and how boxes on a street are being used (trash can, mailbox, etc.). Datasets are now being developed to enable exploration of functional reasoning, particularly for video sequences. Research efforts to integrate functional, appearance and geometric reasoning will yield significantly more meaningful semantic summarization capabilities.

A potentially powerful extension to conventional 3D modeling is the integration of spatial and temporal reasoning to develop 4D scene models. Such 4D models will reconstruct 3D models of both stationary and moving scene entities, place the movers in scene in their 3D coordinates as a function of time, track the movers through the dynamic scene, and enable

immersive scene inspection from novel viewpoints at arbitrary times. Research efforts to develop solutions for these problems needs to address numerous challenges: spatio-temporal image filtering, spatial and temporal camera calibration, 2D and 3D data registration, spatial and temporal uncertainty estimation and propagation, integrated 2D/3D object detection/classification/matching, spatio-temporal context development and application, novel 4D representations, and efficient 4D modeling computation.

Anomaly Detection and Intent Recognition: As semantic summarization capabilities mature, they will not only enable human high-level reasoning, but will also set the stage for automated high-level reasoning algorithms. Sample high level reasoning algorithms include anomaly detection and intent recognition. Anomaly detection develops an understanding for common scene/object attributes so that deviations can be automatically detected. At issue is the need to detect the anomalous behavior in the first place and to avoid high false alarm rates as a side effect. Challenges include the development of normalcy models with possibly limited data and parameterizing the anomaly detection processing with sufficient contextual information to truly separate anomalies from normal behavior.

**Large-scale Visual Recognition:** Investments in addressing the problems listed below are recommended. Robust methods for designing 1000 category object classification and localization for image retrieval purpose are needed. Methods for handling 100 category human action and activity recognition in conjunction with human pose estimation, concurrent action recognition (multiple actions in parallel) are required. At instance level, we will have to solve the long challenges on face recognition, human identification and re-identification from video under reasonable conditions. In the area of scene parsing, and reasoning with commonsense knowledge, methods that an incorporate properties such as physical relation, such as supporting relation between objects are needed. Visual recognition tasks, such as object detection, human pose estimation, vehicle detection, attributes, action recognition, and scene parsing, have been studied in isolation, and must be solved in a joint inference framework. Recognition is only a marginal task of parsing. As a single image or a video clip contains all these visual concepts, all the recognition problems either can be solved together or none of them can be solved. This observation also leads the vision researchers to study the representations and models of commonsense knowledge. This calls for collaborations with language, cognition, and AI. To support joint parsing and reasoning, vision systems must be able to represent and reason with massive commonsense knowledge about images and the physical scenes. Finally, existing vision systems are trained with a pre-defined goal, when this goal changes, say adding a new concept in the hierarchy, the learning process has to start over. To collect the massive commonsense knowledge, the learning process must be lifelong and adaptive to changing goals and utility functions.

**Interrelated Topics:** While the nine themes discussed at the workshop were handled in isolation, it was clear that many topics are interrelated and solving one could lead to better solutions for the other. For example, progress in the field of re-identification is closely tied to biometrics, since biometric features will be essential to any robust re-id strategy. Similarly, there is a two-way relationship between activity recognition and video summarization. Recognizing activities can help re-identification, while establishing the identity of a person can lead to more robust activity labels. Since re-identification involves multiple cameras over a

wide area, suitable summarization can lead to better identification of the regions of interest, and hence, more robust feature extraction. On the other hand, having the identity of a person can yield more informative summarization. Semantic characterization can assist in geo-location, where feature selection and matching algorithms can leverage semantic content understanding. Semantic characterization can also benefit activity understanding, which relies on object detection and attribution produced by semantic summarization. Information from activity understanding results as well as human-in-the-loop interactions can improve the algorithms for semantic characterization. Long-term research activities should focus on exploiting these interrelationships.

**Datasets and Evaluations:** While systematic efforts for data collections and performance evaluations have become routine in many areas discussed in the report, there are several steps that can be taken to further improve this enterprise. First, technology partnerships between members of the research community and those with operational systems in controlled settings where instrumented data collection maybe carried out must be fostered. Second, plans for how to manage privacy concerns relative to data collection and dissemination must be worked out. For example, consider data collection at sites where prior agreement for use of data to support research is a reasonable request likely to be granted. Third, at the planning stages account for ways that both open and sequestered data of similar nature may be acquired and used in both open and more controlled formal (i.e. sequestered) evaluations. Fourth, invest significant resources into open architectures that support research and development, including "blind R&D" methods on sequestered data, as well as technology transfer in a continuous fashion. To be successful, such software must be of value to both researchers and those evaluating technology. When that condition is met, the benefits accrued to both sets of users is dramatic and the pace of technology deployment could be accelerated.

### 1.2.2 Organization of the Report

The report is organized in terms of nine chapters that summarize the panel reports for each of the topic areas mentioned above. Each chapter provides a statement of the problem, relevant work and presents the panelist views on solved, nearly solved problems and problems that need long-term investments. The chapter on data and performance characterization discusses the positive outcomes of data collection and performance evaluation endeavors undertaken recently in many areas considered here and lays out a vision how such efforts should be pursued in the future.

# 2 Video Summarization Panel Report[1]

The problem of generating a shortened summary given a long video has attracted significant attention especially over the past few years. Summarization is a potential enabler for several human-in-the-loop decision systems, where hard-pressed and time-constrained operators have to scan several hours of footage to arrive at critical decisions. City-wide surveillance

[1]by Pavan Turaga, Anuj Srivastava, Jason Thorton, and Shmuel Peleg

in the event of a security threat, surveillance at airports, forensic analysis of multiple video-feeds after an incident, all require pre-filtering of the humongous amounts of video data in an effective manner. Several video abstraction systems have been proposed and good recent surveys with systematic classification of various approaches can be found in [1, 2, 3, 4, 5, 6].

An overview of the summarization pipeline and a simpler categorization is given in Figure 1. We consider three parts of the overall pipeline (a) Triage, (b) Representation and metrics, (c) User presentation and evaluation. We discuss solved, short-term challenges, and long-term problems in terms of these three sub-problems.

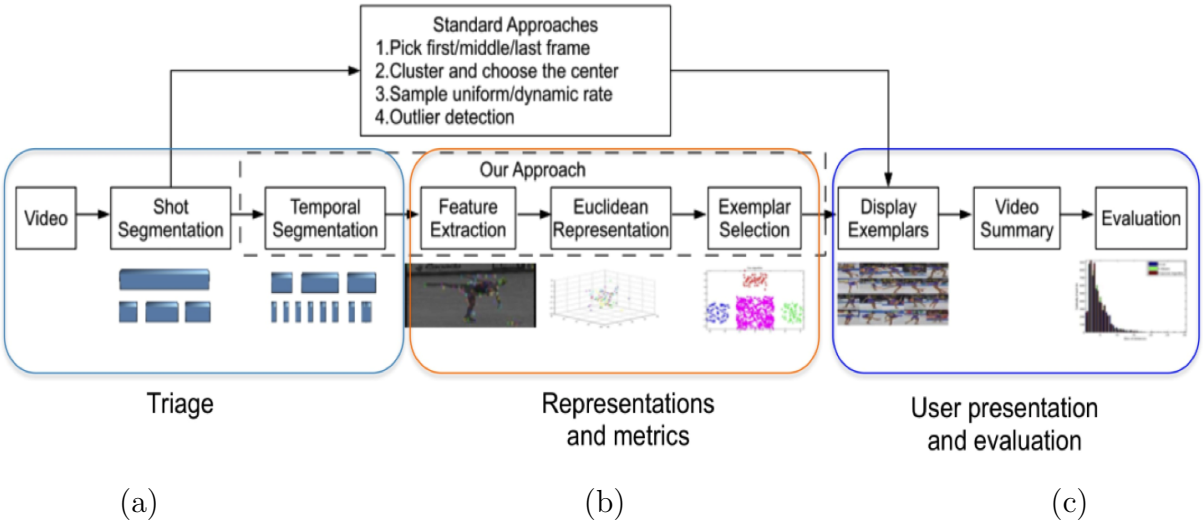



Figure 1: An overview of a typical summarization pipeline, and identification of three core questions (a) Triage, (b) Representation and metrics, (c) User presentation and evaluation.

## 2.1 Solved Problems

### 2.1.1 Triage

Triage is the process of very quickly eliminating irrelevant portions of a video using very simple criteria, while allow later more sophisticated processes to process the reduced set. The most common form of triage has involved detection of shots. A shot is a sequence of frames within a continuous capture-period that has a coherent common trait, such as a common background. The transition between two consecutive shots is termed as the shot boundary. These shot boundaries are then used to detect the temporal span of each shot. Thereafter, exemplars within shots are selected. Initial success using simple techniques such as the random sampling was found to sufficient for simple videos with clean backgrounds, and high quality data. These approaches were motivated from and tuned to genres such as movies and news-videos where the use of shots is an integral part of the video capturing process. In these better understood genres of newscasts and movies, the presence of shots provides cues to interesting content. Thus, simple techniques such as preserving the shot-boundaries worked reasonably well in such genres to provide an overview. Examples of this

approach include [7, 8] who use shot boundary detectors and then use the first/middle/last frames from each shot as the summary. Similarly, [9] choose the first frame of each shot as a key frame, and then if the difference between the subsequent frames and the latest key frame exceeds a threshold, the current frame is also chosen as a key frame. However, this class of approaches does not work well for unconstrained consumer video, such as those found on YouTube, where the notion of shot is not built into the capture process. Thus, a shot-based structure during summarization is solved when shots are natural during acquisition; otherwise there are significant open questions to be considered.

### 2.1.2 Representation and Metrics

In order for summarization to be effective, one needs ways to select informative exemplars from all the available frames or video segments. Classical approaches to summarization consider video features to be points in a high dimensional vector-space, onto which a variety of statistical summarization techniques could be imported. The standard approach is to employ clustering methods. Various image-based features such as color, motion, etc. have been employed to obtain the vectorial representation. Some representative approaches that fall in this category include [10, 11], where the approach uses clustering on the frames within each shot. The frame closest to the center of the largest cluster is selected as the key frame for that shot. This approach does not capture within shot variations of non-stationary shots well, as only one key frame was chosen per shot. [12] propose another clustering technique where all video frames are clustered into a variable number of clusters. Cluster validity analysis is then performed to determine the optimal number of clusters. Similarly, activity specific features are extracted over small segments of videos, and each segment is represented as a collection of features derived from a histogram as in [13] or as a linear dynamical system [14]. The methods presented in [15, 16] have also focused on generating summaries for domain-specific videos where special features using the domain knowledge can be employed. Given a long video, these clustering approaches proceed in an unsupervised fashion to extract summaries. Often times clustering approaches produce redundancy in the sense that the retained cluster centers often are close to high-density regions in the feature space. This corresponds to picking multiple examples from the generally non-informative space of typical actions or concepts. To remedy this issue [17] augment the clustering cost function with a diversity term to obtain representative centers that have higher visual diversity. This was shown to provide better quality summaries for unconstrained videos where simpler pre-processing techniques are not sufficient.

### 2.1.3 User Presentation and Evaluation

Significant research has also been done on finding novel ways of presenting and visualizing obtained summary. One of the most commonly used methods to present summaries is via storyboards. Another approach is through mosaicking [18, 19, 20], which tries to present most of the pixel intensity variations that are spread out over several minutes by piecing together a large mosaic.

Another class of approaches collapses the regions of motion into a smaller spatio-temporal volume. In [21], effective visualization was achieved by first detecting tubes of moving ob-

jects in the video; then collapsing the detected tubes into a shorter coherent video. Infact, this approach was even deployed at FBI according the Professor Shmuel Peleg. An example frame from such a summary is shown in Figure 2. This approach essentially displays all the moving objects in the scene with no special regard to what activity is being performed by the object. A content-aware resizing approach is taken in [22], where seams of low-gradient are successively removed from a video. The resulting video then represents high-gradient information in the video. This video retargeting approach is a useful way of compressing irrelevant spatial information, but it is not clear if this would extend to the temporal dimension as well.


Figure 2: : Example of video synopsis [21], that detects tubes of moving objects and presents a coherent video of the collapsed tubes. Temporality is not necessarily preserved. This technology was acquired by FBI according to Prof. Shmuel Peleg.

## 2.2 Nearly Solved Problems

### 2.2.1 Triage

How does one perform triage of video data when there is no simplified model of video acquisition? What we mean by a simplified model during acquisition includes shot-based acquisition such as in newscast and movies. Triage based on medium- to high-level concept detectors is a possible way forward. The actors/actions of interest are hardly alone in scenes that generally include dynamic backgrounds, interfering objects, partial occlusions, and activity interruptions. There is also a need to register detections across cameras, viewpoints and time stamps. One needs reliable detectors that provide accurate feature estimation for relevant objects despite complex environments. Given progress in machine learning and detection

algorithms, this appears to be a medium-term challenge. A new class of videos that has been recently increasing in popularity is ego-centric video, acquired by wearable cameras. This type of sensing mechanism makes almost every part of the summarization problem hard, and especially triage. Classical approaches based on shots no longer viable, calling for semantic methods for triage. Early work in this area can be found in [23]. Incorporation of user-feedback into the triaging process is also a potentially fruitful approach.

### 2.2.2 Representation and Metrics

The general idea in video summarization is to extract relevant features from data sequences, and provide spatiotemporal summaries of these features, while enabling efficient scene interpretations. In the object-oriented view of the problem, the features are targeted towards characterizing objects and their actions/activities in the observed scenes. While this approach seems intuitive, efficient, there are many challenges to accomplishing these goals.

Novel Feature Space Definitions: Different situations require different features to solve inference problems. Currently there is a lack of specialized discriminative yet efficient set of features (representations, metrics, and statistical models) that allow characterization of object-oriented video data. An emerging theme in object characterization is that relevant features are naturally constrained to be nonlinear and one needs techniques from Riemannian geometry to analyze them. For instance, a statistical analysis of skeleton data in Kinect RGB-D data streams can be efficiently performed using shape manifolds. While shape, texture, and motion are recognized features, especially when searching for individual objects and their actions, there is a need to develop more sophisticated features for characterizing complex scenes involving interactions of numerous objects.

Pose and Rate-Invariance: Another challenge is the omnipresent gap between training and test datasets that results in poor generalizations of activity recognition algorithms to novel test situations. This requires tools that are either invariant or relatively immune to certain nuisance variability (pose, execution rate, sensor registration, etc.) between test and training. In the context of shape analysis of objects (contours, skeletons, etc.), this implies invariance to a larger class – affine and projective – of transformations, beyond the current shape models that handle only similarity transformations. Similarly, when analyzing temporal evolutions of shapes for the purpose of activity classification, it is important to be invariant to the rate of executions. These goals require novel mathematical representations, Riemannian metrics, and inference algorithms that help achieve the desired invariances.

### 2.2.3 Presentation and Evaluation

Characterizing computational bottlenecks: Creating summaries is still a very computationally intensive task, and there have not yet been any systematic studies about computational benchmarks. Consider summarizing a year-long video feed. Even if the video could be processed in real-time, we are looking at a years worth of processing time! Of course, triage can significantly cut down the required time, but still one is faced with a fairly unacceptable amount of processing time. Further, what are the fundamental limits of presenting a years worth of data in, lets say, 5 minutes? Are there fundamental bounds, in the same flavor as found in information theory, which can provide guidance on these issues? These questions

can form the basis of fruitful future work in the medium-term.

Summarization as an evaluation tool: We envision that summarization techniques can be used as tools for evaluating the performance of other tasks. For example, in the near future it is conceivable that lay-users or expert analysts will want to Test-drive an object detector before purchasing it. In such a scenario, how does one allow a user to get a quick sense of how well the detector generalizes to their specific requirement, as op- posed to interpreting the ROC curves of detectors on certain test-beds when the test-beds themselves may or may not be reflective of a users intended deployment scenario? Early work in this area can be found in [24], who use a mix of summarization techniques with group-testing approaches, to allow an end-user to quickly evaluate an object detector (see Figure 3). Summarization can infact be used for evaluating a range of high-level semantic concept detectors and this approach will prove valuable in bridging the oft-stated gap between the performances of algorithms on test-beds, versus their performance as judged by domain experts.

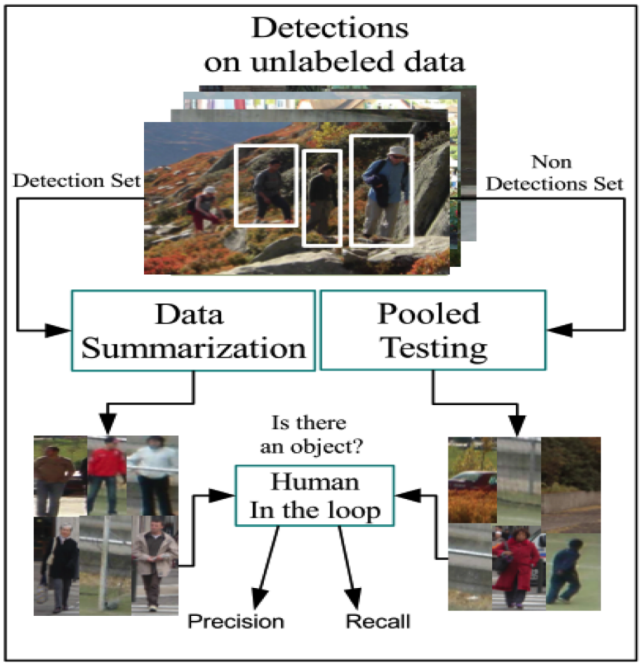



Figure 3: Test-driving a detector via summarization. Estimating the performance of an object detector using summarization techniques with a human in the loop [24].

## 2.3 Problems that Need Long-term Investments

### 2.3.1 Triage

In the long-term, triaging should ideally be based on a user-specified set of rules, that need to be computationally interpreted and translated to image/video operations, rather than work with classical tools which are generally user-agnostic. How can we translate a user-specified

model into an effective triaging platform? This will potentially require a multi-disciplinary approach that bridge image analysis with natural language models.

Further, the notion of video quality needs to be more broadly brought into the triaging approaches. What do we mean by quality is often domain specific, which can range from poor spatial resolution of objects of interest, to noise and jitter, to poor illumination conditions. How can triaging be performed in a manner that is robust to varying quality factors?

### 2.3.2 Representation and Metrics

Perhaps the most difficult issue in video data analysis is to derive high-level interpretations of activities in complex, realistic scenes. Multiple actors, intermittent observations, structured clutter, occlusions, and other complex phenomena often characterize realistic scenes. These factors make the task of scene interpretation quite difficult. In particular, the detection, organization, and understanding of cues and events in real-world scenarios are difficult tasks. One needs to organize information associated with detected objects, and their space-time characterization, into interpretable descriptions. Current efforts borrow ideas from natural language processing, context-free grammars, scene ontologies, and other rule-based patterns, that help organize inferences in complex scenes. Another recent idea is to utilize Grenanders pattern theory that is ideally suited for organizing diverse information extracted by different low-level algorithms, and ontological constraints into Bayesian interpretations of complex scenes. In this theory, individual items such as objects, their actions and interactions are represented by generators and bond connections that allows for discovering statistical relationships between objects. Together these generators and bonds form configurations that provide high-probability interpretations of activities evolving in videos.

It is especially desirable to enhance classification performance of automated systems by using joint, cross-modal statistical inferences. The use of pattern theory allows bridging of gaps between raw signals and high-level, domain-dependent semantics, and helps discovers large groups of audio-visual events likely to represent the same underlying event.

### 2.3.3 User Presentation and Evaluation

Current visualization techniques are mostly limited to visualizing a single camera stream. How can we visualize summaries that come from multiple viewpoints? Perhaps even city-wide camera networks. Perhaps we need to consider 2.5D summarization techniques that in addition to summarizing 2D spatial video can bring into consideration overlapping views, and qualitative geometric models of wide area scenes. This might even mean that summarization research needs to take into consideration novel display modes, beyond just the typical flat-screen display devices. City-wide video summarization can benefits from holographic displays or emerging flexible display hardware.

Finally, the notion of a "good" summary is extremely subjective, as well as context dependent. The panel feels it is not in the interest of the field to evaluate summarization techniques with simple reductionist methods such as ROC curves etc. as found in the detection and classification literature. The quality of summarization needs to studied in-situ, in a deployed environment, with end-users providing qualitative feedback on their experience with these systems. As is the case with most experiential systems, the nature of user-feedback

is often tied in strongly to previous conditioning, the context, and a host of other factors that are not easy to enumerate a priori. These issues can form the basis for long-term research in summarization.

# 3 Visual Analysis and Geo-Localization of Large-Scale Imagery [2]

## 3.1 Introduction

Determining the geo-location of in-the-wild still photos and videos, without the knowledge of metadata such as EXIF and GPS tags, has emerged as an important problem both for its richness of technical challenges and its potential for practical applications in DoD, Intelligence, Industrial and Commercial communities. The timeliness of addressing this problem domain for real world applications is indicated by a number of significant efforts undertaken by various organizations; examples include the IARPA FINDER program [25], the DARPA VMR (Visual Media Reasoning) program [26], and numerous research efforts that are ongoing in academia and industry [27, 28, 29, 30, 31, 32, 33, 34].

The landscape of image and video based geo-localization problem broadly consists of the following three problems:

- Single Image or Video Geo-localization: Given a single photo or a single short video clip, this is the Where is THIS problem. It is assumed that there is no other metadata such as EXIF or GPS tags, or at best metadata is highly uncertain. Examples of the diversity of image content in a single photo on the basis of which geo-location needs to be determined are depicted in Figure 4.


Figure 4: Diversity of image content for the photo geo-localization problem.

- Sequence Geo-Localization: Given a sequence of photos or videos, the problem is to geo-locate the photos and videos without or limited metadata. This problem is similar to the static image geo-location problem but allows the use of temporal information that constrains the trajectory of geo-locations that a camera followed. [31, 34] are two examples of approaches to this problem, one uses video and the other uses sequence of photos to constrain the camera geo-locations.

[2] Jill Crisman, Alexei Efros, John M. Irvine and Harpreet S. Sawhney

- Localization of Non-Overlapping Cameras: The third class of problems in the genre of localization problems includes data obtained from non-overlapping cameras that are spread over a relatively compact region of space and time. For instance, photos and videos provided by social media and surveillance cameras around the time of the Boston bombings in 2013 represent one incarnation of the problem. In order to make sense of 1000s of media clips captured over the area of interest, the problem is to situate the photos and videos in absolute geo-space or relative to each other for human analysts to discern objects, movements and events of interest. An example of this problem is depicted in Figure 5.


Figure 5: The diversity of social media and surveillance camera photos and videos as depicted here for the Boston bombings presents a unique challenge for (geo-localization of all the data for sense making by humans. (Figure courtesy Yaser Sheikh@CMU).

## 3.2 Issues and Challenges in Visual Geo-localization

It is evident from the problems and example imagery depicted above that the metadata-free geo-localization problem is a challenging problem that has ingredients of classic computer vision problems related to visual features, object and entity detection, scene parsing and understanding, and pose estimation and localization. Within this realm, we now highlight the issues and challenges related to image, sequence and heterogeneous imagery localization.

- Geo-localization Uncertainty: In the absolute geo-localization problems the area of uncertainty could be as large as a country or a continent, or potentially even the whole land area of the earth. Coarse metadata or human inputs may reduce the uncertainty to a state, region or a city. However, an ideal system needs to reduce this uncertainty at best to a point on the earth but at least to a small region of the earth.

- Reference Data: Typically geo-localization problems are solved by correlating or aligning photos and videos to a geo-tagged database of reference data. Commonly available reference datasets include 1m. GSD (ground spacing distance) satellite imagery, 30 m. GSD Digital Elevation Maps (DEMs) or Digital Terrain Elevation Data (DTED), and other imagery and data sources of National relevance such as remote sensed data of vegetation types, etc. In the past decade or so, reference data in the form of aerial Lidar 3D point clouds is also becoming available for major cities, and for certain areas of tactical and strategic interest to US Defense and Intelligence operations. Furthermore, Google and Bing StreetView and Oblique imagery also is increasingly becoming available for large areas but with restrictions on use for research and commercial purposes. As is evident from the image examples depicted above, the accuracy and precision of a geo-location solution will critically depend on the accuracy and type of reference data available, as well as the uniqueness and ambiguity of features and layouts seen in the photos and videos. National imagery and terrain data is available for almost all the land cover of the Earth, while visual data as crowd sourced photos and as StreetView imagery are available only for certain popularly visited locations and for cities of commercial value (see Figure 6).

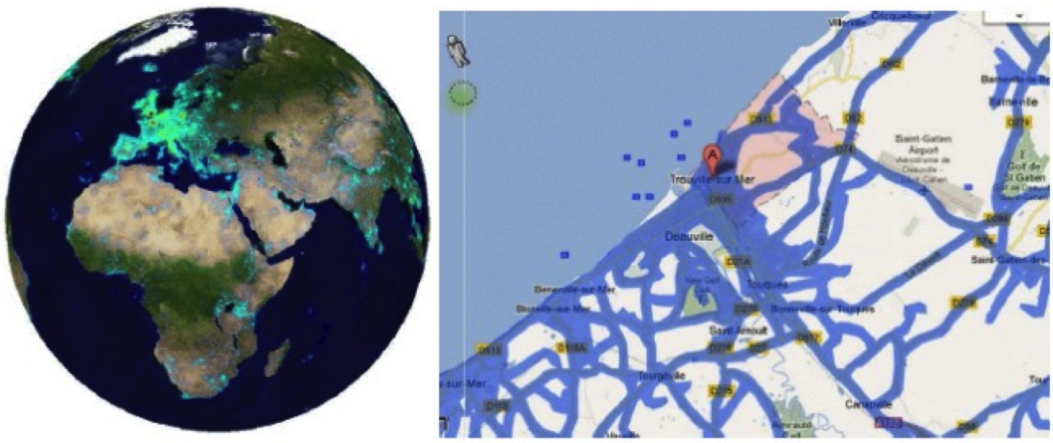


Figure 6: Distribution of crowd photos across the world. Right: StreetView coverage in an urban area of Europe.

- Extreme Viewpoint and Appearance Differences: Since imagery to be geo-located may be taken at arbitrary times and from arbitrary viewpoints, there typically are large differences in appearance and image geometry of features and structures in the captured imagery. For instance, with only satellite imagery as reference data, the problem of finding correspondences between the reference imagery and captured photos is not solvable by direct matching as is traditional in computer vision. Similarly, matching between reference elevation maps and photos is a challenge.

- Diversity of Usable Image Features: Associated with the diversity of image content is the issue of what features can be matched between a reference DB and a captured photo. For instance, for geo-locating street photos in a StreetView database, low-level

features such as SIFT [35, 36] and SURF [37] have been shown to be useful. For geo-locating photos with terrain elevations databases, mid-level features such as skylines have worked well [27]. However, there is a large space of problems in which either the photo is highly ambiguous or the context and content of the image needs to be ascertained for associating with reference databases. In such cases, semantic labeling of the photo and creating a semantic layout [38, 39, 40, 41] may be required to obtain high level descriptions and features that can help in geo-location. In a practical system, typically a user can help by providing appropriate features and probably a strategy for matching.

- Full Automation vs. Semi-Automated Operation: Applications that may require automated reasoning in large collections, geo-location with full automation is a necessity. For applications where a few images are geo-located and an analyst workflow is possible, typically user interaction in the beginning of a session and after a system has returned a set of potential outputs is possible. An interesting problem for semi-automated operation is triage in which a system can query the user, say using a 20-question like interface, to narrow down the area of uncertainty, and also use the user responses to set up appropriate methods for geo-location in terms of features, reference data, and algorithms to be used. Recent work in differentiating fine scale attributes of objects using a 20-question like interface has demonstrated promising results [42].

- Achievable Accuracy: Given the large uncertainty inherent in the geo-location problem, potential solutions will have wide variation in achievable accuracy and precision of solutions depending on type of reference data, photo content and algorithms employed. The spectrum of achievable accuracy and precision can be delineated into three categories:

  - Pinpoint Accuracy: This is the scenario when there is enough information in the image to compute a camera model (intrinsic and/or extrinsic parameters). For instance, a mountain skyline or matched features in a ground imagery database could afford such as computation. Furthermore, a photo may be geo-locatable within a few 100s of meters to a few centimeters. For instance, if there is a distinctive skyline and the reference data can be used to render skylines densely and with high resolution, pinpoint accuracy of a few meters may be achievable. Similarly, with densely sampled StreetView data, street photos are localizable with high pinpoint accuracy.
  - Regional/Scene Accuracy: In scenarios such as the desert scene in Figure 4, without distinctive geometrically localizable features, geo-location methods may only be able to achieve accuracies up to a region on the earths surface. For instance, a system may only be able to say that this is in the Sonoran desert or some location in the Northern Rockies. Or say photos with flat terrain without any distinctive skylines, it may only be possible to obtain an answer such as somewhere in Kansas. Note that if the input uncertainty of geo-location is a whole country or a continent, algorithms that approach regional/scene level accuracy are also quite valuable for high value applications.

– Localization with respect to Scene Elements only: In scenarios such as the one shown in Figure 7, it may only be possible to ascertain a place as a set of visual scene elements. For instance, this seems to be a scene with a tram car from a Czech city, or these cycle rickshaws indicate that this may be Calcutta, etc. This type of localization will in general rely on semantic correlations between visual scene elements from a reference data set to a query photo. In this context, it may be possible to work with relatively spare collections of reference photos with ancillary information coming from visual and text sources such as the open Web, Wikipedia, etc.

Figure 7: A scene with a tram car.

- Performance Characterization: Due to variations in quality of photos, visual content available in a query photo, and the coverage, resolution and quality of the reference data, the output of geo-location needs to be characterized with respect to various statistical measures. Typically, there may be multiple solutions due to ambiguities and the type of algorithms employed. It is desirable that algorithms characterize their performance in terms of precision for uni-modal and multi-modal solutions. This can be in the form of a heat map over the area of interest indicating probability of any given location with respect to the data used and the algorithm employed. In addition to a heat map, to capture the complexity of the solution space, it might also be useful to compute some function of the number of modes and the uncertainty around each mode as a compact measure. One important variable in determining performance is the *quality of a query photo* in terms of blur, contrast, effective resolution and compression artifacts. It may be possible to create a performance prediction model for a given image quality while the other variations are kept constant. Performance characterization can be done for two different purposes. A *Triage* level can inform a user as to what is the expected precision and complexity of a solution without executing an expensive query operation that could take minutes to hours depending on the area of coverage. *Operating Point* level characterization informs a user on the quality of the set of specific solutions obtained.

- Speed and Scalability: The problem of visual geo-location is inherently a large scale

search and matching problem especially for any practical application. A fundamental requirement of any scalable search system is that the complexity of search should at best increase sub-linearly with the size of the database. As a result, direct matching to every database record is not feasible. Feature and attribute indexing to find a short list of potential database records is a viable strategy and has been an active area of research in image retrieval. Also, the database representations may be designed to explicitly address the problem of coarse search without linear matching, and representations that are suited for matching. In general, approaches that perform coarse search in constant or logarithmic time to obtain a short list and subsequently employ fine matching for accuracy and precision have proven viable in practical applications.

- Temporal Constraints: In sequence geo-location problems, in addition to the scene constraints described above, temporal constraints such as travel time priors [34] and ordering constraints [30] can be employed for finding a consistent solution for the full sequence or video rather than frame at a time solutions.

- Self-Localization of Non-Overlapping Cameras: In this problem domain, techniques described above for geo-location of individual cameras can be employed. However, data such as from the Boston bombing provide additional information from spatial correlations across cameras, correlations of movements and actions across cameras, and correlations of entities and activities across cameras. This problem domain has not yet been explored much.

## 3.3 Solved Problems

Solved problem domain can be characterized as one in which pinpoint accuracy is achievable in a semi-automated way with coverage provided by distinctive features in the reference database.

- Skyline based Geo-location: This class of methods computes geo-location of a query photo by matching skylines computed (semi-)automatically in the photo with pre-rendered skylines from a terrain elevation or a Lidar 3D database. Systems based on these methods have already been demonstrated for country sized and larger regions of the earth [27, 43]. These methods are largely applicable to rough, mountainous terrains where distinctive skylines may be available at most locations. Typical approaches render 360 deg. views and extract skylines on a grid of locations that sample the full area of interest at a spacing of a few meters to a kilometer. Skylines can be converted to compact representations using salient features and hash codes. At the time of querying, a skyline is ether computed automatically or through human interaction. The system indexes and matches the query skyline to the stored DB to find the geo-location. Figure 8 illustrates the alignment of a photo to a reference image using skylines.

- Image-based Ground Photo Geo-location: With the robustness of image to image matching obtained by features such as SIFT and SURF, the class of methods that geo-locate by exact matching of such features between a query photo and a database of photos (see Figure 9) is also ready for practical use at least for small to mid-sized

Figure 8: Example of geo-location with skylines by aligning query photo skyline to rendered skyline in a DB. (see http://www.panamintcity.com/exclusives/u2tree.html)

ground level image databases. These methods have been shown to be robust to fairly substantial changes in scale, viewpoint, illumination and image quality [44, 45, 46, 47, 48, 49, 50]. Thorough evaluations have been done both with appearance only and appearance and geometry combined matching. Happily, the low-level features are quite amenable to indexing and hashing and a number of techniques have been developed and evaluated with speed for a large database, and precision with indexing and matching. When multiple overlapping images in crowd sourced data are available for an area of interest, then it is possible to create a 3D representation of the cameras and the structure of the scene in the reference database. This provides further robustness in query photo geo-localization by matching 2D appearance to appearance and 3D geometry representations in the reference database [51, 52, 27].

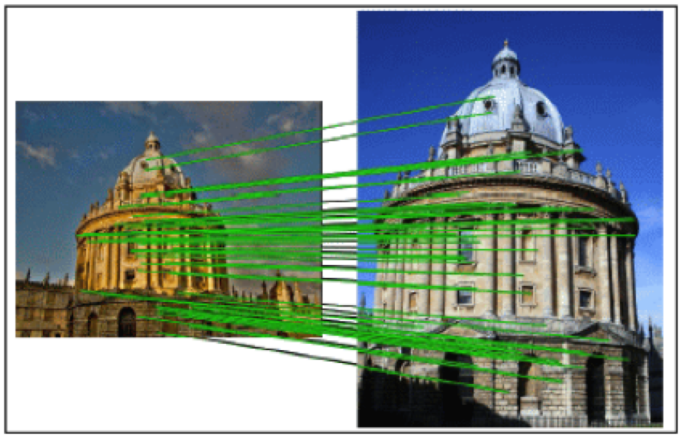


Figure 9: Example of exact matching with SIFT features between two images with different scales and time of capture.

Outstanding challenges in this problem domain include: (1) handling large changes in camera viewpoint, especially rotations in depth; (2) radical changes in illumination such as day/night, haze/clear, etc.; (3) High degree of occlusion from foreground objects and clutter such as foliage; (4) Massive scale localization covering the world.

- Region Location: There is early work on semi-autonomously geo-locating photos using up-to-date land cover classification data of a region of interest (e.g. see the global land cover classification map in Figure 10) [53]. Typically, a user annotates a query photo with classes of region types and also provides a 2D/3D layout of the regions. The methods then use the geo-located land cover classification information to match the types of regions, their extent and topology/geometry. This work is in its early stages but results are encouraging.

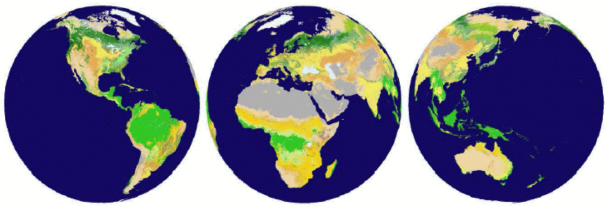


Figure 10: Global land cover classification map.

Within the context of region localization, some work has also been done for the case when exact matching is not possible but global and semi-local characteristics of the visual scene can be used to determine a region of relevance. Myriad image features such as GIST, color distributions, texture distributions, spatial distributions of edge and color distributions, etc. have been used to find a matching region in a database of tagged images using nearest neighbor techniques. An example of this work [29] is shown in Figure 11, a beach scene is located in various regions of the globe.

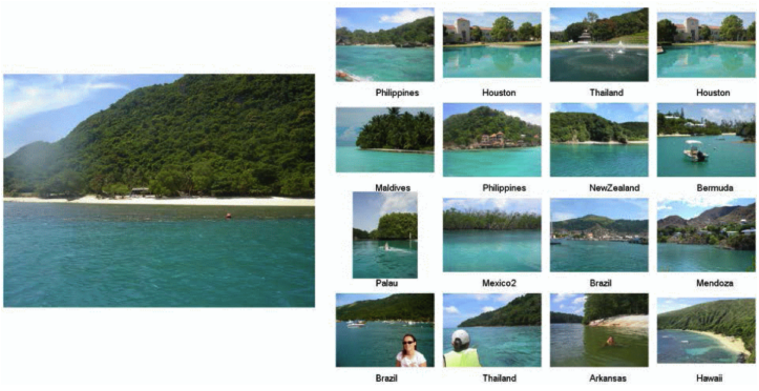



Figure 11: A beach scene located in various parts of the world.

For this scenario of nearest neighbor matching, larger the reference datasets and the more diverse their geographical distributions, the better is the region level precision.

For instance [29] show how the precision increase rapidly with the size of the database (See Figure 12).

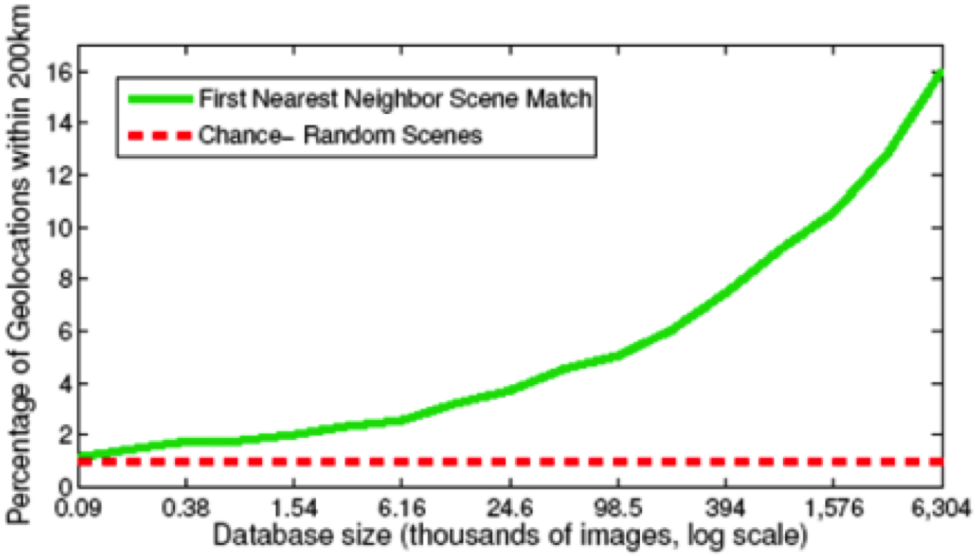



Figure 12: Average precision with respect to dataset size.

## 3.4 Nearly Solved Problems

Moving beyond the currently mature capabilities, achievable capabilities in the next 1-3 years span the gamut of larger scale, more autonomy and performance prediction.

- Full Scale Ground Photo Localization: With advances in feature and image coding, as well as global and local descriptions of images, it will be possible to create systems that use global scale or continent scale photo databases and localize query photos with respect to the stored photos. As currently available, when groups of photos can be used for 3D reconstruction, it will be possible to pinpoint camera location at a large scale otherwise geo-location will be with respect to tagged image locations. It is also conceivable that using the same methodologies, the problem of searching for image patterns in large scale aerial imagery will also have scalable solutions.

- Autonomous Region Localization: We also expect that region localization using land cover classification databases as well as foliage and other databases will be achieved with full autonomy. Classification of image content with respect to certain real world categories of interest is increasingly becoming more precise (for instance, refer to the recent progress in image level classification [54] and object level classification [55] using deep convolutional networks). Also scene segmentation with object and scene models is beginning to show promise. These techniques can be applied to both reference imagery as well as query images on the basis of which categorical and topological matching can be done to arrive at region level localization.

- Sequence and Inter-camera Localization with Visible Landmarks: The algorithmic advances mentioned above can also be applied to sequence and inter-camera localization when landmarks are visible in a sequence or in all the cameras in a set. For this, in addition to appearance and geometry matching within images, cross image sequential temporal constraints or cross-camera topological and geometric constraints will need to be used.

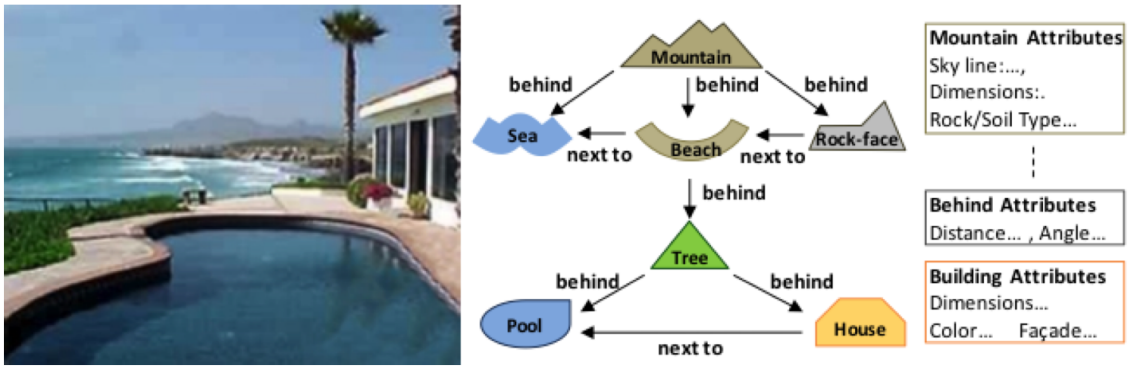
- Performance Prediction: In the next few years, as experience with locating images in large databases grows, techniques for predicting accuracy and precision of geo-location with respect to image content, quality and database content will also be developed. Recent work in the domain of object classification [56] and action recognition [57] is an indicator of some promising directions to follow. Methods for automatically computing image quality will be combined with regression and learning based methods for features-to-location determination to achieve a first level of performance prediction.

## 3.5 Problems that Need Long-term Investments

In the longer term algorithms and systems would (semi-)automatically determine the level of precision achievable for a given geo-location problem and then apply the appropriate methods to get to one of the three precision regimes: visual element location, region location, pinpoint location. When directly matchable, or even regionally matchable attributes are not available in imagery, methods that derive correlations between visual appearance and location using large scale data will be required. In order to derive such correlations, mid-level and high-level (towards semantics) descriptions of images will need to be computed. This is the domain of problems where recognition in the large meets geo-location. In the previously discussed problem domains where exact or approximate matching was possible, potentially machines could have superhuman capability in achieving high levels of precision. However, in the domain of semantics based correlation, we will need to create human like capabilities for reasoning with and matching image attributes. We now describe some of the semantic constraints available in imagery that are indicative of locations.

- Scene Parsing: Parsing of ground imagery as well as aerial/satellite imagery into its semantic components and their 3D topological or geometric relationships will be a big step towards matching such parsed representations for localization. A notional example of such parsing is shown in Figure 13. This process is akin to extracting semantic fingerprints of every location on earth. Early work on 2D / 3D scene parsing includes [38, 39, 40, 41, 58, 59, 60].

  Furthermore, as higher resolutions of National imagery and land classification databases become available, such semantic processing along with precise geo-location of each element in the parse graph will be computed from these databases also. With such scene graph representations derived both from reference databases and query photos, techniques for indexing the appearance and geometry of the underlying elements and their attributes, and fine scale matching for precise localization will need to be developed. It is conceivable that if say at 1m. resolution, the whole earth can be mapped and attributes of each 1m. pixel computed using sensors and processing, then at least




(a) (b) (c)

Figure 13: Semantic Fingerprint Extraction. (a) Ground Image. (b) 3D Parse Graph, and (c) Attributes of Objects and Relationships extracted from the image.

in non-urban areas, localization of ground level views can be solved using the above methodology.

- Urban Localization: For dense urban areas, increasingly 1m or better Lidar data is being created to capture the third dimension (elevation) also. Furthermore, algorithms for parsing large scale Lidar datasets into various physical entities, such as ground plane, buildings, roof types, foliage, road, pavement, poles, etc. [61, 62] is increasingly maturing. As a result, it will be possible to parse large scale urban areas into the physical constituent elements and their appearance and physical attributes. On the query photo end, scene parsing and geometric layout construction techniques, with some human assistance, can help create a 3D layout of surfaces, objects and their constituent elements such as lines and corners. Techniques for matching such descriptions from images to those derived from Lidar could potentially solve the urban localization problem with high levels of precision and accuracy.

- Other Exploitable Constraints: Without the availability of reference datasets mentioned above, regional or visual element localization methods can be used. A variety of constraints can be brought to bear on the problem:

  - Type of Terrain: Large scale, high resolution characterization of terrain types covering the world will enable regional location by matching types of foliage, forests, shrubs, wetland characteristics, agricultural and desert attributes, as well as urban attributes.
  - Visual Element / Attribute Mining: Future work will include learning mid-level and high-level descriptors for visual elements and their attributes that are indicative of locations, cultures, cities, eras, etc. For instance Middle Eastern cities are influenced by the Islamic architecture with its characteristics patterns and constructions, while European cities have characteristics of renaissance and gothic architecture in the patterns of facades, windows, pillar etc. Early work along

these lines [63] indicates that such associations can indeed be learned with lots of data.

- Physics of the Visual World: Cloud patterns, rock formations, terrain characteristics are all weak indicators of regions of the world and the seasons. Exploiting these can involve a combination of modeling of local weather patterns and terrain, as well as learned models of association.
- Patterns of Crops and Vegetation: Similarly vegetation types can be weak to strong indicators of location (See Figure 14). Exploitation of this source of information [64] will involve using text and image sources such as encyclopedias and handbooks to learn associations between locales and vegetation.

Figure 14: Vegetation.

- Correlations of Cultural and Human Artifacts: Man-made artifacts are strong indicators of locales and regions. These artifacts include vehicle types (e.g. autorickshaws), traffic infrastructure (e.g. lane markings and signs), bridges and trains, and other urban and sub-urban infrastructural components.
- Socio-economic Correlations: Presence of commercial logos (e.g. Starbucks), building types and their architecture (e.g high-rise apartment complexes), graffiti and tidiness, etc. are all indicators of socio-economic locale of a community. Extracting these features and attributes from images can be a source of information for visual localization.
- Distance Metrics: When heterogeneous pieces of information need to be combined to get a result, distance metrics for each information source is an important area of research. For instance, in image retrieval, researchers have used weighted distance metrics that reduce the influence of commonly occurring features such as on grass and sky. Also, when scenes are represented as attributed graphs with geometric and topological relationships, it is necessary to learn distance metrics that can measure the similarity between two instances of such representations.
- Putting it All Together: In general the nature of arbitrary photo geo-location problem requires inference of locations by combining weak to strong sources of information. This needs to be done in some applications automatically while others may use human in the loop. Furthermore, generic physical constraints coming from camera models and world models (such as ground plane, horizon, etc.) provide further inputs to an inference engine. Humans can assist in determining what types of features and constraints may be applicable to a given photo

or this may need to be done automatically. Graphical models and information fusion techniques that can work with sources of information with a wide range of uncertainties are required for automated inference of locations and their posterior probability distributions for use by other systems as well as humans. The information extraction and fusion algorithms also need to be structured so that global scales of data can be handled while maintaining acceptable levels of accuracy and precision while keeping the hardware and processing requirements within the limits afforded by various applications. To this end, systems will need to be developed in which each individual feature, attribute and semantic constraint type can be indexed with respect to its location attributes. Search through indices can be the first step in handling scale and speed while finer level information fusion techniques that work on smaller, short-listed database representations can provide the accuracy and precision. This is quite a tall order for extremely in-the-wild geo-location from photos but a pursuit worth going after since it can help make advances in many aspects of computer vision, machine learning and inference, and visual data indexing and searchable representations.

– Sequence Localization and Inter-camera Localization: Most of the techniques mentioned above apply to sequence localization as well. The problem of inter-camera localization may get more attention in the longer term since a lot of the forensic data for events of interest may come from heterogeneous sources of visual information. In this context, the problem of inter-camera localization will be solved by additionally incorporating constraints from scene dynamics in which movements of people, vehicles, crowds and their short-term and longer term identities are used to establish relative localization and topology of a camera network [65].

# 4 Image-based Biometrics [3]

Biometric recognition refers to the automated recognition of individuals based on their anatomical or behavioral traits such as fingerprint, face, iris, gait, and voice. The first scientific paper on automated fingerprint matching was published by Mitchell Trauring in the journal Nature in 1963 [66]. Traurings work was followed by publications outlining automated systems for matching other biometric traits such as voice [67], face [68], signature [69], hand geometry [70] and iris [71]. These pioneering efforts laid the foundation for modern day biometric recognition systems, which are becoming an integral part of several large scale person identification systems around the world today. The identification applications range from law enforcement and civil registry to unlocking mobile devices and entering secure facilities. In this report, we will focus on the three most significant image-based biometrics, face recognition, iris recognition and fingerprint identification.

Work related to biometrics has already had a great deal of real-world impact, perhaps more than any other area of computer vision. Systems for automatic identification of fingerprints have been widely used world-wide for decades. Security systems and access control

[3] Terrance E. Boult, David Jacobs, Anil Jain, David Kriegman, and Marios Savvides

based on fingerprints, iris or face are becoming increasingly popular, in many cases offering greater ease and security than methods based on passwords or keycards. Face detection has achieved wide commercial success; modern digital cameras generally include automatic face detection to improve imaging, including methods that ensure that faces in a photograph are in focus. And automatic face recognition is used commercially in some security settings, and is becoming increasingly popular in systems that allow users to tag personal photographs. Large companies that handle personal photos, such as Apple, Google and Facebook have established significant efforts in face recognition.

Looking to the future, there is increasing demand for improved biometrics. In part this is due to the tremendous growth in the use of digital cameras. Surveillance cameras are so ubiquitous that it is difficult for human operators to monitor them all, while the prevalence of mobile phones with cameras has resulted in a flood of digital images of people. Cybersecurity is an area of growing concern, and existing methods of access control, such as the use of passwords are known to have significant problems (e.g., users often fail to choose secure passwords, or forget their passwords). Security applications create a strong need for better automatic face recognition, as recently highlighted by the search for the Boston Marathon bombers. And while face recognition systems are deployed for auto tagging of personal photos, these systems, while useful, are quite limited in their effectiveness.

Finally, it should be noted that problems in biometrics are connected to some of the most fundamental questions in computer vision and pattern recognition, and advances in biometrics have often had a significant impact on other areas of vision and AI. For example, many new methods that have been developed in statistical pattern recognition for biometrics are broadly applicable. The Viola-Jones face detector [72] developed methods of applying cascades and Adaboost that are applied to a host of other detection problems. Face recognition raises fundamental questions about how variations in lighting and pose can be accounted for in image matching, and progress on these problems has been applied to a range of other recognition problems.

For all these reasons, visual biometrics is an extremely important area of computer vision. Progress in this area can take advantage of a large commercial infrastructure that is poised to incorporate new research ideas into useful products and systems. At the same time, the deep problems (e.g., representation, matching, throughput, spoof detection, template security) posed by biometrics ensure that work in this area can have broad, long-term scientific impact.

Biometric systems are experiencing continuous improvements in performance and usability. Many independent third-party technology evaluations have been conducted primarily by NIST for fingerprint, face, and iris. These NIST evaluations serve as an excellent resource to benchmark the current recognition performance of various biometric systems. In general, the error rates of a biometric system depend on a number of test conditions (See Figure 15). Consequently, the NIST evaluations tend to be quite extensive and include results obtained under a variety of test conditions.

## 4.1 Solved Problems

- Fingerprints Recognition: The Fingerprint Vendor Technology Evaluation (FpVTE) conducted by NIST in 2003 [73] shows that the best commercial fingerprint recognition system can achieve a True Acceptance Rate (TAR) of 99:4% at a False Acceptance Rate

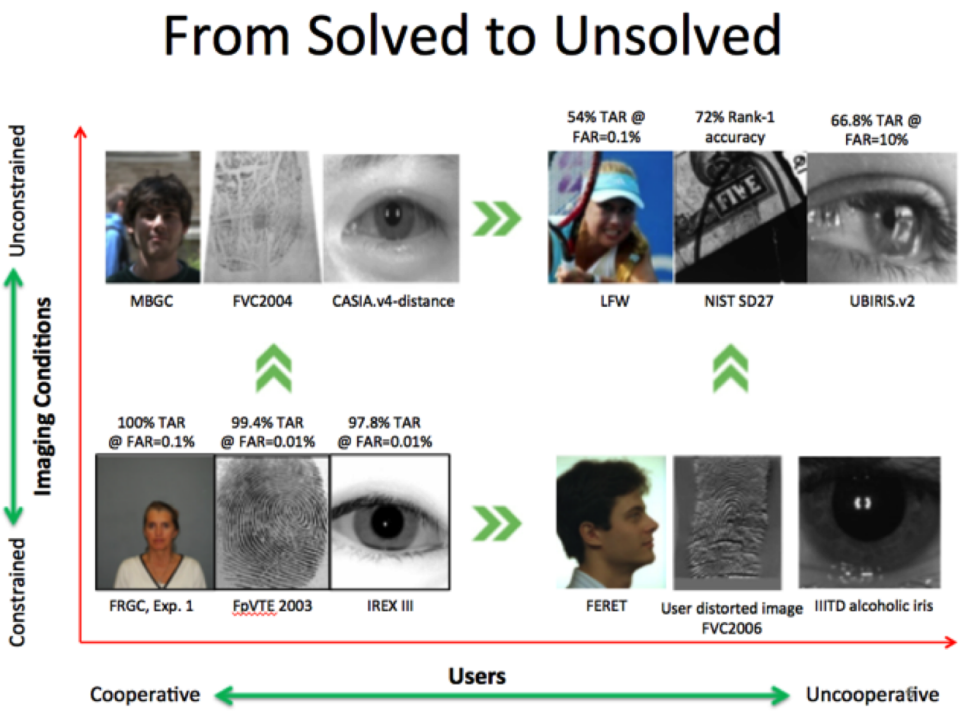



Figure 15: Researchers have explored a wide spectrum of problems. The difficulty increases as acquisition conditions become less constrained, and when subjects are not actively cooperating with the system.

(FAR) of 0:01% for plain-to-plain matching based on fingerprint data collected from various government sources in the United States. While a more recent evaluation for plain-to-plain matching called FpVTE 2012 has been conducted by NIST, the results of this evaluation have not yet been released.

The results of FpVTE 2003 and other such evaluations indicate that the technology for plain-to-plain (as well as rolled-to-rolled) fingerprint matching is fairly mature and very high accuracy can be obtained under typical conditions. However, there may be some scope for improvement in accuracy when the user is uncooperative and provides distorted or partial fingerprint images or if the image quality is very poor due to finger skin conditions. The results of different phases of Evaluation of Latent Fingerprint Technologies (ELFT) conducted by NIST confirm that the problem of latent-to-rolled print matching is inherently more challenging compared to plain-to-plain matching. The best rank-1 accuracy obtained in ELFT-EFS Phase 2 was only 63:4%. The largest public-domain latent fingerprint database is the NIST Special Database-27 (NIST SD-27) and the best reported rank-1 accuracy on this database is 72%. These numbers clearly show that the problem of fully automated processing and matching of latent prints to rolled impressions or other latent prints is still far from being solved.

- Face Recognition: Evaluating the state-of-the-art in face recognition is more complex because of the large scope for variability in face images due to a number of factors including aging, pose, expression, and illumination. The NIST Face Recognition Vendor Test (FRVT) 2012 indicates that face recognition systems can achieve a TAR (True Acceptance Rate) of approximately 96% at a FAR (False Acceptance Rate) of 0.1% when matching mug shots of the face (frontal face images obtained under a controlled environment at the time a suspect is booked at the police station). When presented with face images obtained during the visa application process, the TAR improves to nearly 99% at the same FAR of 0.1%. This is because face images for visa processing have more stringent guidelines on illumination, background, and occlusion. While the above results are impressive, they are applicable only to a small number of applications where good quality face images can be captured from cooperative subjects in constrained environments. However, the face images captured covertly in surveillance applications tend to exhibit more intra-class variations. A reasonable indicator of the performance under such challenging conditions is the accuracy of various face recognition algorithms on the NIST Special Database-32, which is also known as the Multiple Encounter Dataset (MEDS). This database contains face images exhibiting relatively large intra-class variations such as pose and illumination changes, compared to mug shots.

  Another widely used benchmark for unconstrained face recognition is the Labeled Faces in the Wild (LFW) dataset [74]. LFW contains 13,233 images of 5,749 individuals acquired from Yahoo News in 2002. Results have been reported from 38 commercial and academic methods. Since 2007, matching accuracy on this dataset has climbed from an equal error rate (equal number of false positives and false negatives in a verification task) of 72% to over 97% in the most recent results [75]. These news photos are taken under less constrained conditions than those used in many prior

tests, with significant variations in lighting, pose and facial expression, and frequent occlusions. Still, news photos are generally taken by professional photographers, and selected by editors. Also, LFW images were restricted to include only faces that could be detected by current automatic face detection algorithms. For these reasons, even this data set is much more constrained than images that might be found in personal photo collections, or that might be captured by video surveillance systems.

- Iris Recognition: One of the largest independent technology evaluations of iris recognition is the NIST IREX III evaluation. The database used in this test contained approximately 6.1 million iris images acquired from nearly 4.3 million eyes. Among the 95 different algorithms considered in this evaluation, the best algorithm had a false negative identification rate of approximately 2.5% when a single eye was used per person and the threshold was set such that there were no more than 25 false positives in every 1,013 iris comparisons. This ability to operate at a very low probability of a false match is one of the key advantages of iris recognition. It was observed that pupil dilation and constriction had a significant impact on the recognition accuracy. Iris images that were not recognized correctly during IREX III evaluation are of very poor quality, mainly because the users did not interact correctly with the iris sensor.

- Perception: Because of the "CSI-effect" there are many in both the government and the public that grossly overestimate the abilities of identification technologies. A difficult, but far less technical issue for biometrics/identity technology is that of perception, which has two dimensions. The first dimension is that the CSI-effect results in many people considering biometrics and other identity issues to be "solved" and hence they no longer consider them important directions for "research topics". This is unfortunate since the solutions to identity applications that are often portrayed in television far exceed state of the art technology, but clearly have widespread potential. The second dimension is that the medias portrayal of identity technologies fuels fears of big-brother spying, which reduces public support for research in these areas. While the privacy issue is clearly of concern, government policies to address them are not yet in place and funding opportunities for biometric system security are lacking.

## 4.2 Nearly Solved Problems

In both the short and long term, much of the emphasis in visual biometrics is going to be on handling unconstrained, natural imagery. The difficulty of unconstrained recognition problems depends in part on the degree to which the conditions of the probe and gallery images match, or to which the gallery images must be extrapolated to account for new conditions. For example, recognizing a subjects face when seen from an arbitrary viewpoint is a challenging problem. However, the problem is easier if one has available a gallery of images of the subject taken from many viewpoints, which can be compared to a new, probe image. The problem is much more difficult if the gallery includes only a single, frontal image of the subject. In the short term, we expect to see more progress on the first set of problems.

There are also considerable variations in the difficulty of unconstrained face recognition, depending on the domain of interest. For example, as mentioned above, photos available on

the internet are in some sense curated for quality. In personal photo collections, quality may be more variable, including more blurry images. But usually, in personal photos, the people who are of interest are captured at good resolution, facing the camera. Images taken from surveillance cameras are even more challenging; subjects generally do not face the camera and resolution may be quite low. Clearly, more rapid progress can be expected on the easier subset of these problems.

Because image-based biometrics is a relatively mature field, with many commercial systems deployed, we expect that any significant research gains can be rapidly incorporated into deployed systems. There is rapid progress being made in improving the performance of biometric systems in unconstrained settings. For example, face recognition systems are showing much improved performance when there is (still somewhat modest) variation in pose, and significant variation in lighting and facial expression. Performance has also drastically improved in related problems, such as face detection and the detection of fiducial points on faces (e.g., the corners of the eyes and mouth, the tip of the nose) in unconstrained settings. While achieving a very high level of performance in completely unconstrained settings is a long-range challenge, in the short term we can expect continuing improvements in performance, with greater accuracy in increasingly varied settings.

## 4.3 Problems that Need Long-term Investments

Highly accurate visual biometrics in unconstrained settings is one of the most challenging problems in computer vision and pattern recognition. Complete solutions to such a problem are beyond the current horizon, but we can expect continuing rapid progress that significantly extends the applicability of current biometric systems.

In the arena of face recognition, significant progress has been made on handling the problems created by variations in viewpoint, lighting and resolution. These are the most well-defined and well-understood challenges of unconstrained face recognition, and it is to be expected that they will be the first problems to be successfully addressed. There are versions of these problems that are extremely difficult, such as recognizing an individual in extremely low-resolution images, or recognizing a profile view of a person when only a frontal view is present in the gallery. But one should expect significant progress on all but the most challenging conditions.

Challenges due to changes in facial expression and the passage of time (aging) are more challenging, and less well understood, because these variations are more difficult to model accurately. Changes in facial expression can be handled to some extent by using robust matching methods that disregard parts of the face that are affected by expression change. But to understand that images of two mouths are from the same person when one is broadly smiling and the other has a neutral expression poses significant long range challenges. Similarly, many of the variations in appearance that occur over time, such as due to aging or weight change, are not currently well understood. The problem is further confounded when one or more of these variations are observed simultaneously in the face image.

- Individuality of Biometric Traits: The concept of quantifying the uniqueness of a biometric trait (in other words, estimating the individuality of a biometric trait) can be

understood from the following simple analogy. Suppose that users in a person recognition system are identified based on a 10-digit personal identification number (PIN). If each user selects an arbitrary PIN in a completely random fashion, the theoretical limit on the number of users who can be uniquely identified by such a system is 10 billion. Similarly, we can say the probability that any two users in such a person recognition system will have the same PIN is 1 in 10 billion. However, it is impossible to achieve this theoretical limit in practice because the users seldom choose a random PIN.

Similar to the case of PIN, it is important to know how many users can be uniquely identified by a biometric system based on a specific biometric trait (e.g., right index fingerprint). This information is often critical when designing large-scale biometric identification systems such as the national ID programs around the world. If a single biometric trait is not sufficient for unique identification, it is essential to know how many traits (multiple fingerprints, multiple irises, face, etc.) would be required to uniquely identify all individuals in a target population. Finally, forensic applications require an estimate of the probability that any two or more individuals may have sufficiently similar biometric samples in a given target population. This is necessary to provide credence to latent fingerprint evidence.

- Biometric Aging: Aging refers to changes in a biometric trait or the corresponding template over a time span, which can potentially impact the accuracy of a biometric matcher. For the sake of clarity, we distinguish between two types of aging: trait aging and template aging. Trait aging refers to the biological change in a trait over a persons lifetime. This change is inevitable and, unlike other types of intra-class variations, cannot be easily controlled by the individual. For example, changes in a persons facial structure and appearance can occur over time due to the effects of biological aging. This can, in turn, impact the accuracy of appearance-based face matchers.

  Template aging, on the other hand, refers to changes in a persons biometric template (i.e., the feature set extracted from the biometric trait) over time. While template aging is certainly related to trait aging, it must be noted that the extraction of invariant features from a biometric trait can mitigate the impact of trait aging on template aging. For example, the appearance of a persons fingerprint is known to vary over time due to age-related as well as occupation related changes in the outer skin, sebaceous gland activity, etc. However, these changes are unlikely to impact the distribution of minutiae points on the fingerprint.

  This problem, however, is much more challenging in the case of face recognition. The human face undergoes very significant changes over time. Some of these changes are due to aging, as the face sags and wrinkles and lines appear. But significant changes can occur even over a very short period of time, as people gain or lose weight or change their eye ware or facial hair. These changes are not yet very well modeled by existing approaches. Further, there is a lack of representations that are invariant to the passage of time for human faces, in the same way that they have been developed for fingerprints.

  One general question that needs to be answered is the following: Can the degree of permanence of a biometric trait/template be computed? In other words, is it possible to measure and predict the degree of change that a certain trait or template is expected

to encounter over an individuals lifetime? An answer to this question would allow for the system to periodically, and systematically, update the biometric template of a user in order to account for age related changes [76].

A second important question is: Can we model the way in which a biometric trait/template changes over time? If we can model these changes, this would facilitate our ability to extract features that are invariant or insensitive to these changes, or perhaps to even predict these expected changes. For example, given a ten-year old photo taken when a subject was forty, can we predict the type of wrinkles that can be found on his or her present day face?

- Unconstrained Sensing Environment: There are some person recognition applications where it is very difficult to impose constraints on how the biometric trait should be acquired. One well-known example is latent fingerprints acquired from crime scenes. The challenges involved in latent fingerprint matching have already been pointed out earlier. One of the major roadblocks in the adoption of iris recognition is the poor usability of iris sensors. Most available iris sensors require the users eye to be in close proximity to the camera and expect the user to remain still during the acquisition process. User acceptance of iris recognition technology can be greatly enhanced if iris sensors can be designed to capture the iris pattern at a distance and when the user is on the move. However, the iris images obtained in this scenario are unlikely to record the texture details on the iris surface with high fidelity and may also exhibit large intra-class variations (e.g., rotation and occlusion). Hence, more robust algorithms are required to process such iris images.

  Unconstrained sensing is a problem of particular importance in face recognition because of the huge and growing number of images and videos, ranging from personal photos to surveillance videos that are taken in unconstrained settings. Most face recognition databases today are constructed under carefully controlled conditions, such as drivers license photographs, passport photographs, criminal or previous offender mug shots, etc. wherein the individual is requested to look at a camera, under uniform lighting, with minimal expression or occlusion artifacts. Real-world image capture in crime and intelligence analysis, however, rarely mirrors these characteristics; they notoriously depict the EUROPIA problems (Expression, Unconstrained Resolution, Occlusions, Pose, Illuminations and Aging). Worse still, these problems are exacerbated by the use of poor imaging devices. The importance of solving this problem, particularly the low resolution challenge was made particularly evident during the events of the Boston marathon bombing in April 2013 (See Figure 16).

  Low resolution images are the result of how far the object of interest is from the camera (standoff), the specifications of the imaging sensor (and its tolerance to noise and low-light capabilities), and the quality of the optical lenses attached to the sensor. The pervasiveness of affordable digital imaging devices, such as cell phone cameras, surveillance cameras, etc., comes as a result of the recent mass production and the shrinking in size of these devices. However, this does not always translate into increased image quality, and in general the visual quality of footage obtained by these devices remains poor. The problem is exacerbated by the ever-shrinking size of the CCD sensor, which


Figure 16: . Surveillance images released by the FBI during the Boston marathon bombing event in April 2013, highlighting the importance of the development of low-resolution face recognition algorithms.

increases the amount of noise in the image, and the use of wide-angle lenses to increase the view angle which in return introduces barrel distortion. Moreover, the popular need for wireless communications with the devices has forced the manufacturers to implement aggressive compression algorithms to increase throughput, but in return, further introduce image artifacts and blurring to the footage. Finally, in surveillance applications, most cameras are placed high up on a wall or on a ceiling to offer them a vantage viewpoint. However, this makes the object of interest too distant from the camera and decreases the number of pixels apportioned to the face.

An unconstrained sensing environment also means that there will be a significant difference in the pose, expression and illumination between the gallery images and a probe image. While each one of these topics has received significant attention in the research community, long-term challenges remain. In particular, while theoretically appealing methods exist for handling each of these variations in isolation, there is little work that effectively synthesizes these results, handling the interaction between all these effects.

Illumination has perhaps received the most attention, and many illumination insensitive image representations have been proposed, including Local Binary Patterns, Gabor jets, direction of image gradient, and various types of filtering to reduce the effects of illumination [77]. These methods produce excellent results on constrained test sets in which only illumination is allowed to vary. However, many of these representations are sensitive to changes in pose or expression, and it is not clear how illumination variation can be best accounted for when all these effects are present together.

Similarly, a number of methods have been developed to handle pose variation [78], including the use of stereo-based matching algorithms, and learning based approaches that determine how pose variations can be factored out of matching. Morphable models

have shown strong results by fitting a 3D model to a 2D image and then virtually rotating the image to a frontal position. All of these methods degrade in performance, however, when expression variation is present in an image.

Facial expression, also, has been handled effectively when it is the only image variation present. It should be said, though, that expression is perhaps the least understood of these three effects, because of the difficulty of accurately modeling the range of possible expression variations. Because of the complex interactions between all these variations, it is rarely the case that a method that is designed to handle just one variation can be effectively integrated into a system that handles fully unconstrained images, such as LFW. Clearly, both more research is needed on each of these effects, and in particular, the interaction between them needs to be better understood.

- System Security and User Privacy: While the main motivation for deploying a biometric system is to protect an application from unauthorized access, there is no guarantee that a biometric system will be completely secure. Just like any other security system, the biometric system may be vulnerable to a number of security threats, which may eventually affect the security of the end application. These security vulnerabilities may lead to adverse consequences such as denial-of-service to legitimate users, intrusion by unauthorized users, repudiation claims by corrupt users, and erosion of user privacy due to function creep.

  While many of the adversarial attacks on a biometric system such as Trojan horse, replay, and man-in-the-middle attacks are common to any authentication system, there are two vulnerabilities that are more specific to biometric systems. One of them is the problem of spoofing, where the biometric sensor is presented with a counterfeit biometric trait that is not obtained from a live person [79, 80, 81]. Spoof detection is a critical requirement, especially in unsupervised applications (e.g., authentication on a smartphone) where the presence of a user is not being monitored. Ongoing research in anti-spoofing [82, 83] is making progress but even for the most well studied issues of face and finger, is still a nascent area. The other major threat is the system security and user privacy issues arising from the leakage of biometric template information due to attacks on the template database. Intentional alteration of biometric traits [84, 85] in order to avoid identification is also an emerging threat in some applications (e.g., international border crossing). It must be emphasized that biometric system security and user privacy concerns are important public perception issues that can potentially derail the success of a biometric system deployment unless they are addressed comprehensively.

  A related issue is the "biometrics dilemma [86] that wider deployment of biometrics systems for security will degrade their potential for security because the underlying data cannot be revoked when compromised. These issues have both strong security implications as well as personal privacy issues.

  While there is a growing body of research trying to address this issue via template protection [87, 88] and/or Biocryptographic technologies [86], these technologies are still less accurate than non-protected techniques and hence research is needed to improve their accuracy. While programs such as NSTIC had the potential to address such issues, it did not, and no existing program is addressing these issues. In addition, some of

these approaches offer new functional capabilities for multi-factor authentication with privacy while directly supporting key management, which could be transformative for cyber-security applications and cyber-resilience.

- Context in Face Recognition: Most work on face recognition has attempted to identify a face using a cropped window of the image containing a single face. However, in many cases, using a larger context may assist in recognition. For example, in organizing personal photos, a trivial constraint is that the same person cannot appear more than once in the same photo. More generally, some people may tend to appear together (your aunt and uncle), so the identity of one person may provide evidence about the identity of another.

- Attributes and Descriptive Biometrics: An important new direction in vision-based research is the use of describable visual attribute, commonly just called attributes. This work, pioneered by [89, 90] builds classifiers that map low-level image features into terms humans might use to describe aspects of the scene. Attributes are of interest because they are a natural method for humans to communicate their understanding/descriptions of people and objects. The use of describable visual attributes for identification has been around since antiquity, being used in ancient Greece, but until recently has not been the focus of work by researchers in computer vision.

  When attribute technology was introduced (2008/2009), they provided a significant improvement in face verification and text-based searching for faces. Since that time, hundreds of other papers have cited or built on that work. Attributes have now been extended into detection of scars/marks/tattoos [91] and general scene attributes [92] and fine-grained object classification [93]. They have been demonstrated to support search with near real-time complex queries about faces being issued over millions of images on a laptop [94].

  In the 3-5 year range, attribute technology has a strong potential for direct impact. One example of future potential is connecting computer vision techniques with human intelligence and textual sources of "big data". While standard biometrics can be used in some settings, they are not well suited to matching with witness descriptions or human intelligence. While many companies and research groups can match faces, there are times when the problem is really about matching faces based on textual descriptions. One obvious application is to allow Google-like searches over faces. As shown in Figure 17, descriptive biometrics have even greater potential beyond searching they can provide the connective glue to allow identity intelligence (biometrics and video surveillance) to be linked to other forms of intelligence. Obvious uses include intelligence and forensic work. Descriptive biometrics also has the ability to begin to allow deduplication and linking of disparate intelligence databases. To develop such applications needs research and development not just in attributes, which need better probabilistic representations, but in how to connect and reason about contextual factors to produce probabilistic representations that fuse the attributes with existing intelligence information and witness descriptions.

  A related approach to allow users to more easily navigate large collections of face images is to build systems that can cluster faces. For example, consider video surveillance data

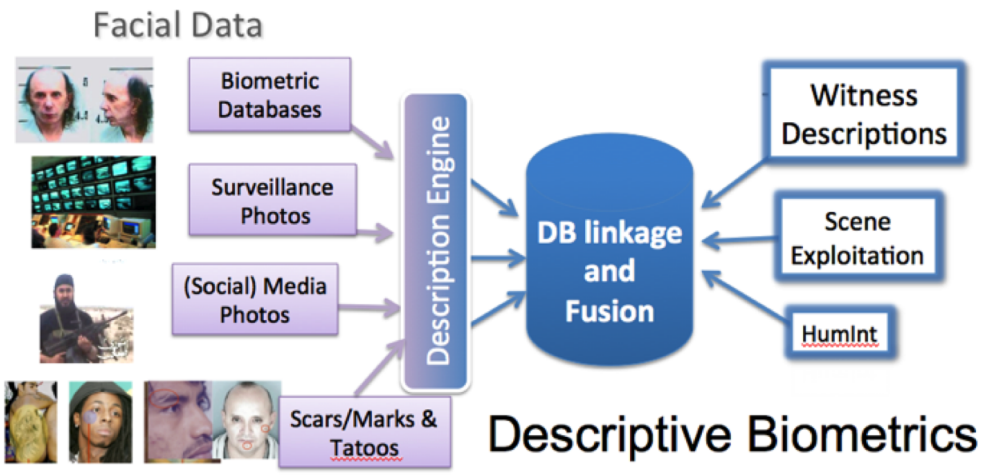



Figure 17: Attributes will enable descriptive biometrics to provide a link between biometrics and human intelligence and witness descriptions enabling new types of searches and better utilization of biometric data.

in airports. An analyst may want to search through the surveillance data to find if a specific person was present in the airport on a particular day. If we can cluster the face images into groups of similar appearance, it may be significantly easier for the analyst to browse them to detect the presence of that person.

Longer term and more open research issues for attribute technology are integration of identity attributes information with semantic labeling and could be integrated with geolocation and other technologies to be able to answer key questions of "Who What and Where".

- Big Data for Face Recognition: Recent advances in technology have made it possible to build very large labeled image sets. Many approaches to face recognition and detection already take advantage of data sets containing tens of thousands of images, and recent results use a proprietary dataset of several million labeled faces [75]. However, it may become possible in the near future to build even larger image sets. New research is needed both to determine how to build these image sets efficiently and to understand how such image sets can be used most effectively.

# 5 Humans in the Loop[4]

## 5.1 Introduction

Human in the Loop (HIL) research includes all the aspects of image and video analysis (IVA) in which people are actively involved. This encompasses three main types of research activity. The first is research in the design of IVA systems to be used by humans. A typical example is

[4]Shih-Fu Chang, Kristen Grauman, Lina Karam, Deepak Khosla, and Nuno Vasconcelos

a system that will aid an intelligence analyst in some IVA task that involves large quantities of imagery. This could range from inspecting large collections of satellite imagery to spot targets of interest, to the inspection of hours of video for forensic analysis after a terrorist attack. The aftermath of the recent Boston bombing incident is a good example. The goal is to build IVA systems that leverage image and action understanding capabilities, to produce intuitive interfaces for retrieval, browsing, and summarization of image and video data. This could critically diminish the time required to identify suspects of criminal activity, targets of intelligence analysis, etc.

A second area of HIL research is the design of systems that leverage the work of humans to improve IVA performance. This is mostly an attempt to deal with the fact that IVA is a very difficult problem. The dominant research trend within this area is crowd sourcing. This could range from systems that ask humans to solve IVA tasks, to systems that use humans to produce training data for IVA systems. For example, when a preeminent Microsoft researcher (Jim Gray) disappeared after a solo trip with his sail boat outside San Francisco Bay, a group of Microsoft researchers were able to convince the Digital Global Satellite to make a run over the area. They then placed thousands of images on the Amazon cloud service and created tasks for reviewing these images. Within two days, 6,000 volunteer Amazon Turkers had inspected more than 100,000 images [95]. Unfortunately this was not sufficient to locate the missing sailboat. A similar episode has occurred recently with the lost Malaysian Airlines airplane.

Finally, a third major area of HIL research is the design of systems that embed humans in IVA. The dominant research trend in this area is "first-person" IVA. This consists of the analysis of images and video collected by wearable cameras, such as the recently introduced Google glass platform. Such systems are likely to become ubiquitous in the near future, both in the realm of civil society and military forces. In this context, IVA can aid the users of these systems by augmenting the context of their physical experience, by retrieving information that augments the physical scene, or provides some context for it.

There are four main components to the HIL problem. In this section, we will describe them in the context of the image retrieval problem, but similar descriptions can be given for all the other problems mentioned above. The first is the component of visual analysis. This consists of the computer vision operations used to process all images and video. The classical paradigm for content-based image retrieval is query by visual example (QBVE) [96]. QBVE systems retrieve images using strict visual matching, ranking database images by similarity to a user-provided query image. A QBVE system extracts a visual signature from the query, compares this signature to those previously computed for the images in the database, and returns the closest matches. The left side of Figure 18 presents an example of the images retrieved by these types of systems. There are many ways to compose image signatures or evaluate their similarity. While early solutions relied on very simple image-processing techniques, such as matching histograms of image colors, modern systems rely on more sophisticated representations and aim for provably optimal retrieval performance. This is the visual analysis component of the retrieval problem.

However, simple visual similarity is not enough for satisfactory retrieval performance, since there are many queries for which visual similarity does not correlate strongly with human similarity judgments. This can lead to a semantic gap between user and machine. Figure 19 presents a subtle example of how people frequently discard strong visual cues in

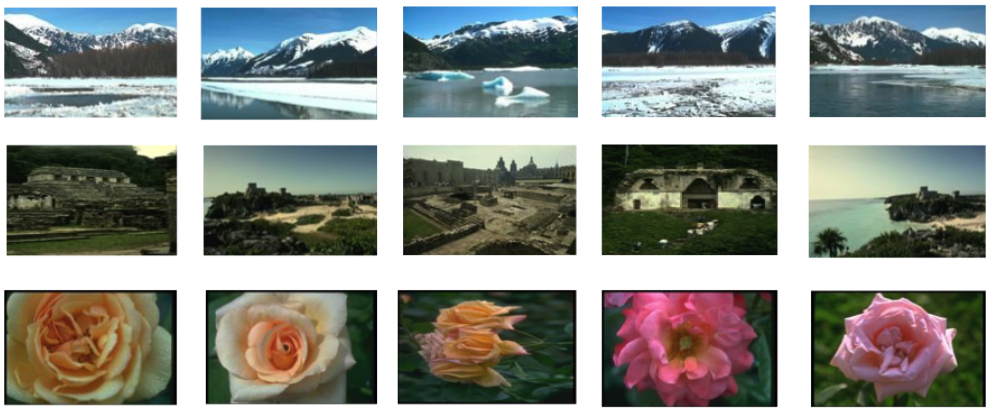


Figure 18: Each row shows the top matches in an image database to the query image shown on the left.

their similarity judgments. The "train" query contains a predominant arch-like structure that, from a strictly visual standpoint, makes the query highly compatible with concepts such as "bridge" or "arch". A QBVE system will return as top matches images like the four shown, three of which indeed contain bridges or arch-like structures. However, people expect images of trains among the retrieved results and assign little probability to alternative interpretations, such as bridge or arch. They seem to decide first that the image is about trains, and then use "trainness" as the dimension that determines image similarity. Whether other trains are visually similar to what the query depicts - for example, in terms of colors, shape, or size - is relatively unimportant. This mismatch between similarity judgments can leave users convinced that the system doesnt get it. It motivates a second important component of the HIL problem, the component of perceptual modeling. Since HIL systems revolve around humans, they usually require some ability to model, or replicate human perception. For example, the use of image similarity measures that reflect those used by humans [97, 98]. In the absence of this it is difficult to guarantee a meaningful user experience.

Obviously, the use of perceptual models does not guarantee a satisfactory user experience. A third major component of the HIL problem is that of interactivity. This consists of the design of mechanisms that allow the human to interact with the IVA system so as to accomplish a certain task. In the retrieval scenario, interactivity is usually implemented with resort to relevance feedback algorithms. These are algorithms which take input from the user, so as to improve the image search. For example, the user may click on a few images that are similar to the target image and some others which are not a good match. The system then updates the retrieved set to account for the relevance feedback [99]. This type of interactivity is usually connected to on-line learning algorithms [100]. The retrieval system may also perform active learning [101], so as to jointly achieve the goals of retrieval accuracy and exploration, i.e. returning a mix of images that are likely to be what the user wants and a set of images that are maximally informative of the remainder of the database. By selecting these images, the user can explore the contents of the database. This is particularly important when users are not really sure of what they are looking for.

The final component of HIL system is that of human computer interface (HCI) design.

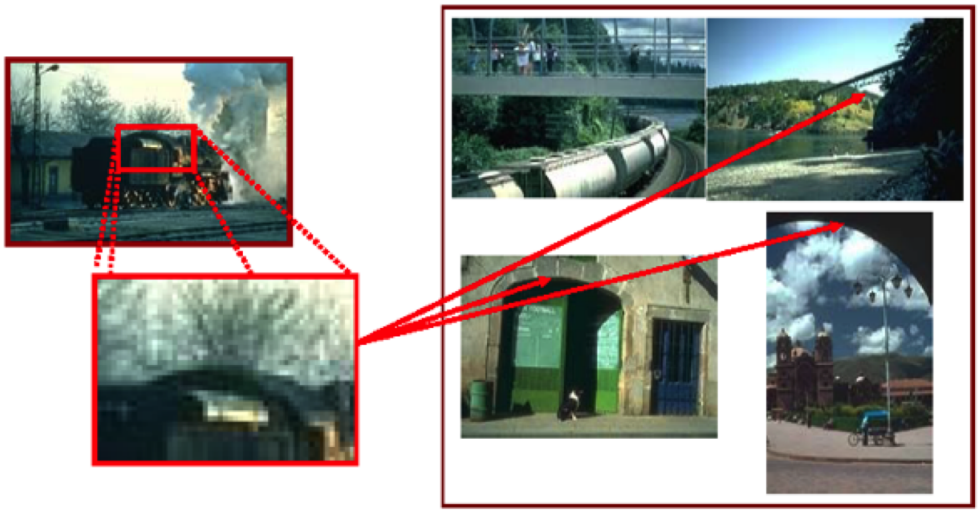


Figure 19: A query image containing a train and the top four database matches according to a QBVE system.

This follows naturally from the need to present information to the user, receive user feedback, etc. In the retrieval example, HCI issues include visualization techniques for the exploration of the image database, organization of information presented to the user, different types of interfaces for collecting user feedback (e.g. eye trackers or even brain machine interfaces), etc.

## 5.2 Solved Problems

Like most IVA problems, HIL is a very difficult goal, currently far from well understood. Hence, most of the research in the area has still not produced solutions that are robust enough for industrial deployment. There are nevertheless some success stories. One example is work on image retrieval based on keypoint matching [36]. Several such systems have been developed by both academia and industry, most notably the Google Goggles app. These systems work quite well for the retrieval of images of identical objects under different poses, especially if the objects are planar (logos, book covers, etc.) or very distinct. They are typically used to match a picture taken with a personal camera (e.g. cellphone or Google glass) to a database of products, landmarks, etc., so as to obtain additional information about an object. Figure 20 shows an example of how Google Goggles can be used to find out information about a famous landmark. These systems cannot, however, perform object class recognition, e.g. match an image of a chair to an image of another chair of different visual appearance. In this domain, there has been significant success in the recognition of a few object classes, which are important for many applications. The most notable among these are faces [72], which can now be recognized by virtually any wearable camera, cell-phone, etc. and car or pedestrian detection, which will become ubiquitous in vehicles during the next decade [102]. However, the general problem of recognizing classes of objects is still wide open.

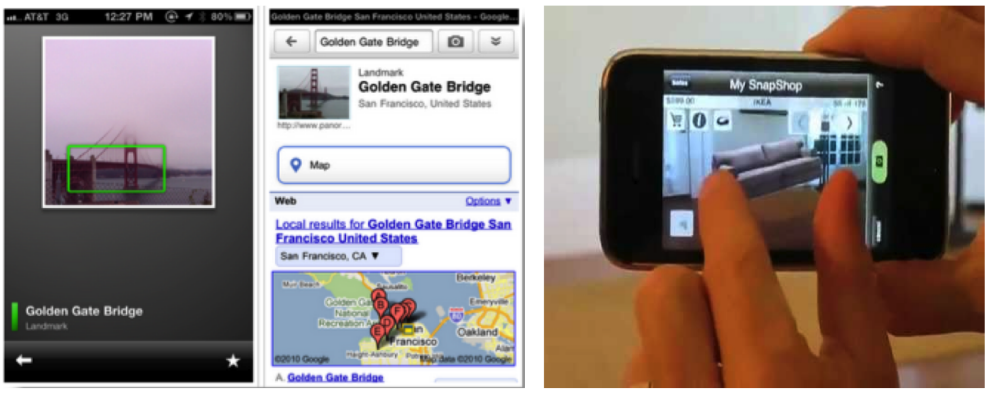



Figure 20: Left the Google goggles system. Right: Adding virtual furniture to a living room with My SnapShop.

Another success story is the combination of geometric computer vision techniques and first person video, to enable some form of augmented reality. These systems allow the superposition of 3D computer graphics on images acquired by a wearable camera, cell phone, etc. This can be used to place objects in new environments or augment scenes with all types of information. Qualcomms augmented reality Vuforia platform [103] is a popular tool for developing these types of applications. In the area of crowd-sourcing, there have also been substantial advances in infrastructure, e.g. by the introduction of platforms like Amazon Turk, which are now in routine use. These systems have already achieved significant breakthroughs, such as the solution of an open problem on the structure of an AIDS virus by gamers on fold.it [104], or the mapping of the retinal connectome by the EyeWire project [105]. A number of technological advances, such as the introduction of CAPTCHAs [**?**] for human identification, were critical for the development of these systems. Finally, many systems with an HIL component have been built in the arena of IVA for surveillance. These include systems for the detection of intrusions, systems for retrieving events in surveillance video, etc. In general, these systems can perform simple operations (such as virtual trip-wires) very well, but fail for more complex queries (e.g. subjects planting explosives, etc.).

## 5.3 Near-term Solvable Problems

It is difficult to predict what will be solved in the next five years. Instead, in this section we will attempt to summarize a number of HIL areas that are currently the subject of substantial attention and could produce significant advances in this time period. The presentation is organized along the components of the HIL problem.

In the area of visual analysis, the current buzzword is deep learning [106, 54]. It seems likely that the next five years will see substantial computer vision work in this area and some of it will be leveraged to improve HIL systems. Classical problems like semantic segmentation, object recognition, image hashing, multi-object tracking, action recognition, and person identification are also likely to be the subject of non-trivial improvements. It is unlikely, however, that any of them will be solved in five years. Another predictable trend is a more pronounced emphasis on scalability. Many groups are now adopting large

datasets, like ImageNet [107] for image classification or TREC MED for action recognition, containing millions of images and large numbers of classes, which will be the norm in the next five years. This will motivate work on scalability issues, platforms for large scale vision, fast implementations of vision algorithms, etc. Some of the HIL problems, e.g. retrieval systems for intelligence analysis or complexity constrained video processing on Google glass are likely to benefit from this.

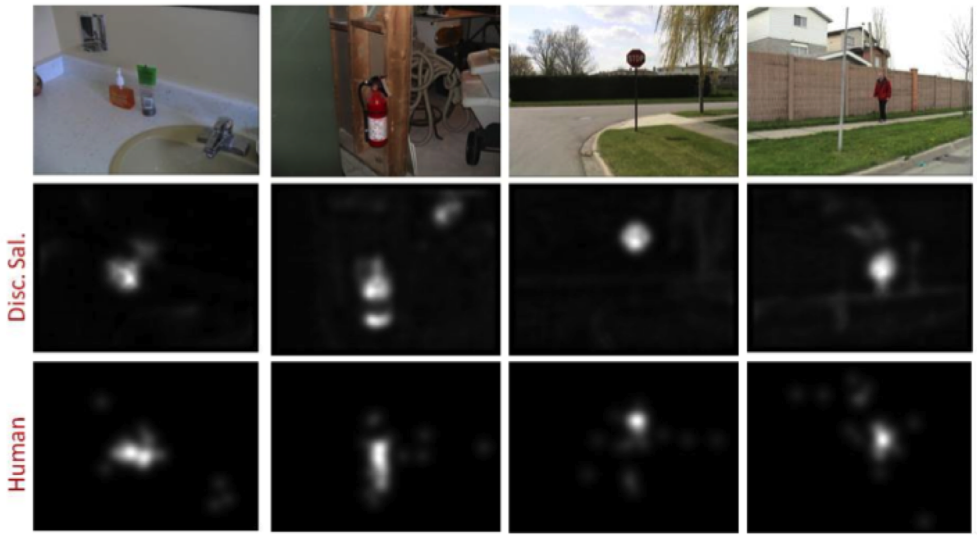



Figure 21: Example predictions of a saliency model that replicates many psychophysical traits of human attention and neurophysiologic traits of the visual cortex.

In the area of perceptual modeling, we believe that substantial advances will be achieved in the area of attention modeling. This is a topic of importance for all components of HIL. In IVA systems used by humans (e.g. retrieval) attention models can be used to emphasize perceptually salient image and video regions, enabling a better match between computer vision and perception. In IVA systems that leverage humans (e.g. crowd sourcing) they can be used to help focus the attention of human workers on the tasks at hand. In embedded IVA systems (e.g. first person video) they can be used to predict where people are or should be looking at. The past five years have already produced substantial advances in attention modeling [108]. Many algorithms have been proposed to model saliency, the stimulus driven component of the attention system. Figure 21 presents some example predictions of a saliency model that replicates many psychophysical traits of human attention and neurophysiologic traits of the visual cortex [109]. This type of "bottom-up saliency is now fairly well understood and there are several models that achieve human-level performance. However, the prediction of salient image locations is not sufficient for many applications, since these locations can frequently lack semantics (they can be simply edges or blobs). Recently, there has been a shift to the problem of object saliency, where the goal is to detect salient objects [110]. Figure 22 presents examples of a state of the art results in this area [111]. Performance is already good for simple scenes, where the background is not very cluttered. The next five years should produce improvements in robustness to clutter.

Figure 22: State of the art results in object saliency.

## 5.4 Problems that Need Long-term Investments

There are several long-term questions for the HIL problem. Many of these originate in open questions in the area of visual analysis, which is itself a widely open problem at this point. However, in many cases, these questions are also central to other areas of HIL, such as perceptual modeling or interactivity. Hence, in this section, we abandon the area centric format used above and concentrate more on the individual problems. Whenever appropriate, we make connections to the different areas.

- Cross modality: Humans think along many dimensions, fusing visual, auditory, memory based (priors), and tactual information, among others. Most people can imagine how a certain animal looks like, how a certain type of food smells, etc. All this information is combined to perform perceptual tasks, with each modality providing constraints to the others. Central to this process is the establishment of cross-modal representations, i.e. representations which abolish barriers across modalities, bringing information from all the modalities into a universal coordinate frame. This type of abstraction is what allows us to choose the image that best illustrates a blurb of text. In artificial perception these representations have not received much attention. Vision researchers work on vision, text researchers on text, audio researchers on sound, etc. Different modalities are sometimes combined in multimedia research, but usually at a superficial level, e.g. by simply concatenating feature vectors from different modalities. This is a multimodal representation, but not necessarily cross-modal, since it may not necessarily establish bridges between the different modalities. To be truly cross-modal, a representation must support inference even when some of the modalities are hidden or unavailable. While some initial steps have been done along this direction, they tend to involve limited amounts of multimodal information, e.g. tasks involving images and text labels or brief text captions [112]. Little work has been performed on tasks involving multiple modalities, where each modality is represented by copious amounts of information. Support for this type of inference could lead to major breakthroughs for the HIL problem. For example systems that allow the use of an image to query a text database. Or, as illustrated in Figure 23, vision systems that use hidden text to im-

prove the generalization ability of operations such as scene classification or recognition. An early solution to this type of problem is discussed in [113].

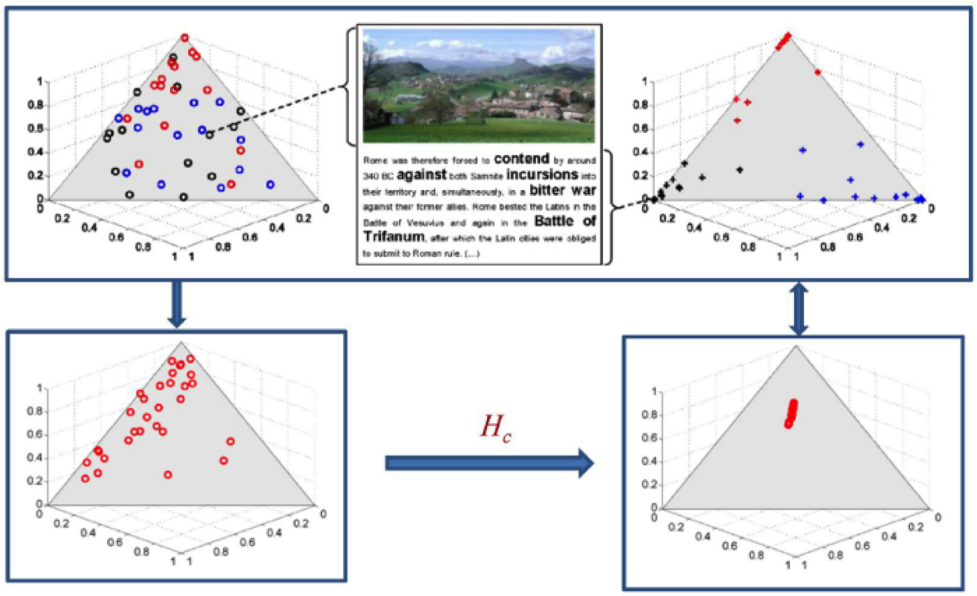



Figure 23: Using text to improve image classification. Top: features extracted from the image (left) and text (right) components of a multimodal dataset. Bottom: the text information is used to regularize visual classifiers, by learning a transformation (Hc) that maps the noisy image features to the cleaner text features. Outside of the training set, this transformation is used to denoise the image classifier outputs, enabling improved image classification.

- Adaptation and transfer across domains: Humans can quickly transfer knowledge from one task to the next [114], from one modality to another, etc. This is a central requirement for the human ability for zero-shot learning, i.e. learning new concepts from a very limited number of examples. Most users expect smart systems to exhibit this type of abilities. For example, an intelligence analyst would expect an image search system to learn from the patterns of interaction between the two. Similar types of searches should provide similar results, but take less time as the system knows more about its user. However, current HIL systems provide very little support for this type of functionality. This problem is present in all areas of HIL, from vision systems that cannot transfer information across camera views, to interfaces that cannot transfer information across user sessions, to perceptual models that cannot learn the commonalities and intricacies of different users. Over the last decade, the machine learning community has started to formalize some of these problems, in the form of the domain adaptation [115, 116, 117], model adaptation [118, 119, 120], multitask [121] or transfer learning [122, 123] problems. Example questions include how to combine a model, learned under setting Y, with a small amount of data collected in setting X [124, 125, 126], how to map a feature space Y into a feature space X [127, 128], etc. While there has been some progress in all these problems, both the theoretical and algorithmic foundations

are still in their infancy. It is also only now that enough computation and storage are becoming available for researchers to worry about problems such as dataset bias [129], i.e. how well an algorithm learned under certain training conditions generalizes to others.

- Semantic representations: Humans abstract information into semantic representations involving abstract concepts. For example, they think of a "park bench" and "bean bag" objects as similar because both are objects to sit in. When they make mistakes, these tend not to be random but to follow these semantics: things that are semantically similar are more likely to be confused than things that are semantically unrelated. When interacting with "smart" systems, users can much more easily tolerate these types of mistakes. For example, an image retrieval system that confuses a bird with an airplane is much less frustrating to interact with than one that confuses the airplane with a fish. A semantic representation is a representation in a space with semantics. This is usually obtained by defining a vocabulary, training a number of classifiers to detect images, sounds, text, etc. of each vocabulary class, and running all the data through these classifiers. An illustration of the process is given in Figure 24. Semantic representations have various interesting properties. First, they eliminate barriers across modalities. After the semantic mapping, all data is represented as concept vectors, independently of whether it was extracted from text, images, or sound. Hence, semantic representations facilitate transfer leaning. Second, they enable the design of similarity functions that more closely mimic those used by humans. For example, concepts can be mapped into taxonomies and these taxonomies used to inform perceptually consistent measures of distance between images or sounds. This leads to systems that make more tolerable mistakes. Third, they make it easier for humans to provide feedback. While it is intuitive for a nave user to tell a system (which confuses birds and airplanes) that airplanes do not usually have a beak, it is much less clear what feedback should be given to a system that confuses airplanes with fish.

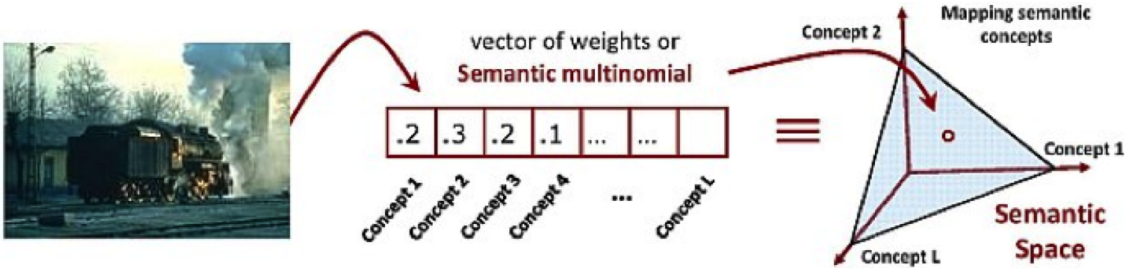



Figure 24: Semantic image representation. Images are represented by vectors of probability of containing various visual concepts. This leads to semantic feature spaces, where each axis is associated with a semantic concept.

While semantic representations have long been used in multimedia and more recently in vision [130, 98, 131, 132, 133, 134, 135, 136], there are still a number of hard open questions in this field. One obvious limitation is that most semantic representations are flat, i.e. assume that all concepts have equal semantic level. While a few efforts have

attempted to use taxonomies, these are usually hard-coded, e.g. based on Wordnet. On the other hand, human taxonomies are a mix of common sense knowledge and user experience, acquired through years of interaction with the world. There is currently very little ability to learn a taxonomy, or adapt an existing taxonomy to an user, e.g. based on the patterns of interaction with an IVA system. This is, in fact, another instance of the transfer learning problem. There has also been little work in trying to define semantic metric structures that truly mimic those used in human similarity judgments, or to design cross-modal systems that exploit these structures to transfer knowledge across information modalities. These, and most other problems in the area of semantic representations, are likely to become more central in the coming decades, as large quantities of storage and computing become available, enabling the routine design of semantic representations with thousands of concepts.

- Biologically inspired models: it could be said that IVA is a mantle of a large number of very difficult problems. To be fully useful, IVA systems will eventually have to solve problems like object recognition, tracking, and segmentation, while explaining away distractors such as occlusions, complex backgrounds, and sophisticated object interactions, with robustness to scale, viewpoint, and lighting variability, and accommodating for variable user feedback, large scale searches, complexity constraints, etc. While progress has been made in each of these sub-domains, this progress has been achieved mostly at the cost of specialization. Different sub-communities have evolved to address each of these different sub-problems, using different formalisms and mathematical tools. It could be said that, today, no one is really trying to solve the IVA problem, but just slices of it. Unfortunately, it also seems to be the case that these problems cannot be solved in isolation, e.g. image segmentation requires object recognition, which requires understanding of scene context, which requires geometric modeling, which requires modeling of image correspondences, and so on. Even worse, the architectures of choice of the different sub-communities are not necessarily compatible and certainly not synergistic. This makes it very hard to contemplate the design of universal vision systems.

  If it were not for the fact that our brains can solve the problem, most vision researchers would probably declare this goal as impossible. This merits the question of whether there are many ways to solve vision other than those found by biology. It could be that the solution of all these problems, with a modular architecture, which is scalable and robust, requires the use of very specific image representations, which coincidentally map all these complicated sub-domains of vision into a unique problem that can be solved with simple and modular circuits. The paradox of biological vision is that the vision questions indeed seem to be easy to solve, with most of the effort devoted to guaranteeing that unrelated constraints, such energy efficiency, are met. For example, which computer vision researcher would think of designing a camera that keeps saccading through a scene, multiple times per second? For biology, this appears not to be a problem: it is clearly more important to save the energy spent in head movements than to simplify the vision problem that the brain must solve to align the information collected by these saccades.

It could thus be argued that the study of biological vision systems, the development of computational models that explain the functionality of these systems, and the translation of these models into computer vision algorithms that solve multiple vision tasks, should be the subject of much heavier emphasis in computer vision than they are today. While the recent excitement about deep learning has some of these characteristics, current deep learning models are far from realistic models of biological computing. Critical information processing components, such as 1) the feedback from higher cognition necessary to implement "top-down" vision routines [137] or 2) different types of normalization [138], are simply not accounted for. A better understanding of these components, as well as of the principles that guide biological computation, could thus prove critical to enable universal vision systems.

- Interactive learning: the current paradigms for learning from and responding to user feedback are based on the notion of risk minimization. This is an estimate of average system performance, which works well for problems such as classification and regression and is quite popular in machine learning. Most on-line learning or active learning algorithms, on which HIL systems depend, are based on these principles. However, in the HIL context, optimal performance on average is not necessarily satisfying. An intelligence analyst is not necessarily looking for a system that works well on average, but instead for a system that works well for his/her particular intelligence analysis query. This can lead to substantial frustration with IVA systems. For example, most image retrieval systems work well for a certain percentage of the queries. However, when this is not the case, it may be quite difficult to steer the system to the target imagery. Even worse, the progress towards these targets can be quite unintuitive. It is not infrequent for the response of the system to user feedback to be a solution worse than that which motivated the feedback. Sometimes, after a few iterations of progress, the system will abruptly get off track and present a solution that makes little sense. This is partly because the system may make incorrect inferences, e.g. because it relies on faulty computer vision, but also partly because its optimization criterion (average performance) does not guarantee good performance for any particular session of user interaction, just for the average across all sessions. Addressing this problem requires alternative learning paradigms, e.g. optimization of worst-case instead of average performance, which have been so far poorly explored in the machine learning literature.

- Weakly-supervised learning: Machine learning research has overwhelming addressed the problems of supervised (classification, regression) and unsupervised (clustering, deep network) learning. Yet, in the HIL context, virtually all problems lie in between these two extremes, or what is usually called weakly supervised learning. This is because supervision tends to be tedious and cannot ever be acquired in full detail. For example, in a crowd-sourcing system for labeling images, human labelers are typically asked to provide a few words that best describe an image or video. Since an image is worth a 1000 words, this usually leaves out much of the image content. In result, while the images are labeled, the labels cannot be fully trusted. Typically, the absence of a label does not imply that the image lacks the associated visual concept, simply

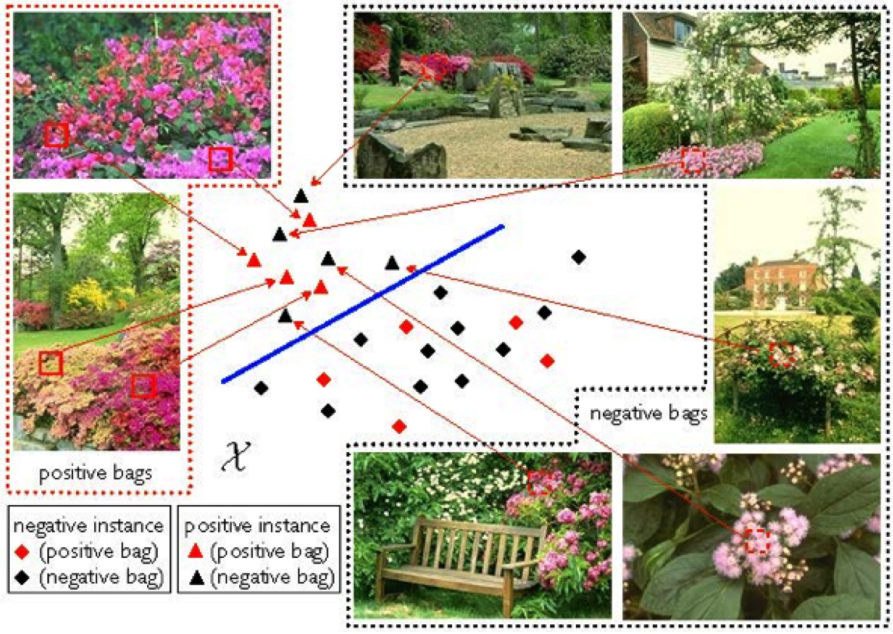



Figure 25: Weakly supervised learning of the concept "flowers". Image of "flowers" (positive bags) contain many flowerless regions (negative instances) and flowers (positive instances) appear in images that do not have the flowers label (negative bags).

that the labeler did not consider that concept to be important enough. Similarly, the presence of a label does not imply that the image is all about that label. This is illustrated in Figure 25. When HIL labels are handed out to a supervised learning algorithm, performance tends to be quite low. These algorithms are simply not robust enough to label noise. On the other hand, it makes little sense to simply ignore the labels and rely on unsupervised learning, which usually does not have very high performance to start with. The right paradigm for this type of problems is multiple instance learning [139]. Under this paradigm, examples (denoted as instances), are grouped into bags, and a label is attached to each bag. A bag that contains at least one positive example is considered a positive bag, otherwise it is negative. A classifier is finally designed to classify bags, rather than individual examples. While there are a number of multiple instance learning algorithms available [140, 141, 142, 112, 143], the theoretical principles have not been thoroughly studied, and there are very few guarantees for generalization, noise robustness, etc. Advances in the theory of weakly-supervised learning and the consequent development of learning algorithms with better performance guarantees are thus of great importance for the HIL area.

# 6 Person Re-Identification

## 6.1 Introduction

Person re-identification (re-id for short) is a fundamental task in automated video surveillance. Given an image/video of a person taken from one camera, re-identification is the process of identifying the person from images/videos taken from a different camera. Re-identification is indispensable in establishing consistent labeling across multiple cameras or even within the same camera to re-establish disconnected or lost tracks. Re-identification is a difficult problem because of the visual ambiguity and spatiotemporal uncertainty in a persons appearance across different cameras. These difficulties are often compounded by low-resolution images or poor quality video feeds with large amounts of unrelated information in them that does not aid the re-identification problem. Algorithmically, re-identification can be defined as a process of establishing correspondence between images of a person taken from different cameras. It is used to determine whether instances captured by different cameras belong to the same person; in other words, assign a stable ID to different instances of the person.

## 6.2 State of the Art

In the last few years, the problem of re-identifying persons across multiple non-overlapping cameras has received increasing attention (see **Table 1** listing the main contributions in the literature). The community has commonly adopted three different kinds of approaches:

1. discriminative signatures based methods,
2. metric learning based methods,
3. transformation learning based methods.

A multidimensional taxonomy and categorization of the person re-identification algorithms can be obtained in recent review papers [144, 145]. In the recent book [146], one can find the current trends in re-identification, also from a multi-sensory perspective. In the following, we provide a brief review of the existing re-identification methods.

- Discriminative signature-based methods: Bag-of-words representations of color, shape and texture features have been the most common choice for discriminative signature based methods. In [147, 148, 149, 150] multiple local features are used to compute discriminative signatures for each person using multiple images. In [151], frames are used to build a collaborative representation that best approximates the query frames. A part-based spatio-temporal model based on HS color histograms and representative colors is used in [152]. The persons body is divided into stable body parts using HOG based body part detectors. The color histograms are extracted for each body part and are combined into an active color model. Representative colors are also extracted from each body part and combined over multiple images by clustering to generate representative meta colors. The active color model and representative meta colors are used as the appearance descriptor and similarity is computed as a weighted

sum over the two features. The same model is extended in [153] to include facial features from low-resolution face images in order to assist re-identification. In [154], Mean Riemannian Covariance (MRC) patches extracted from a particular individual are used in a boosting scheme. In [155], re-identification is performed by matching shape descriptors of color distributions projected in the log-chromaticity space from different targets. In [156], an adjacency constrained patch matching strategy based on an unsupervised salient feature learning framework is used to improve re-identification accuracy. In [157], the BiCov image representation relying on the combination of Biologically Inspired Features (BIF) and covariance descriptors is used to compute the similarity between person images.

- Metric-learning based methods: According to [158], in a metric learning framework a set of training data is used to learn an optimal non-Euclidean metric which minimizes the distance between features of pairs of true matches while maximizing the same between pairs of wrong matches. In [159] the re-identification problem is formulated as a local distance comparison problem by introducing an energy-based loss function that measures the similarity between appearance instances. In [160], a metric just based on equivalence constraints is learned. In [161], a relaxation of the positivity constraint of the Mahalanobis metric introduced by defining two different classes for pairs of true matches and pairs of wrong matches reduced the computational cost yet gave competitive performance. Some of the recent works try to improve the metric learning performance by excluding well separable examples and solving an eigenvalue problem [162], by giving less importance to unfamiliar matches in a large margin nearest neighbor framework [158], by learning multiple metrics specific to different candidate sets in a transfer learning set up [163] and by exploiting sparse pairwise similarity/dissimilarity constraints [164]. In [165], a metric learning procedure based on Local Fisher Discriminant Analysis is applied after a PCA dimensionality reduction step to maintain redundancy in color-space. In [166], re-identification is performed by measuring cosine similarity between the gallery and the probe descriptors which, in turn, are constructed by measuring similarity with the reference data in a Regularized Canonical Correlation Analysis (RCCA) subspace. In [167], re-id is cast as a multi-view semi-supervised multi-class recognition problem [168], where each class corresponds to the identity of one individual. In this setting, each feature is associated with a component of a vector-valued function in a vector-valued Reproducing Kernel Hilbert Space, and combined by a fusion mechanism in this framework. It does not require inter-camera image pairs of the same person, and can work with only a single labeled image per person  a situation often encountered in real scenarios. For a more thorough review of metric learning approaches, the interested readers are directed to two survey papers [169, 170] on this subject. These methods, address the target re-identification problem relying on the discriminating power of the proposed signatures and proper distance metrics.

- Learning feature transformations: One of the early works of finding the transformation of features between cameras was proposed in [171]. A Brightness Transfer Function (BTF) between the appearance features is computed by finding the optimal path in the feature correlation matrix. A similar approach is proposed in [172] where a learned

subspace of the computed BTFs is used to match the targets. An incremental learning framework to model linear color variations between cameras is proposed in [173]. Both [172] and [173] learned space-time probabilities of moving targets between cameras and used them as cues for association. However, transition time information may be unreliable if camera FoVs are significantly non-overlapping. In [174] a sparse color information preserving Cumulative BTF (CBTF) is learned from training examples collected from a pair of camera views. In [175] a Weighted Brightness Transfer Function (WBTF) that assigns unequal weights to observations based on how close they are to test observations is proposed. In [176] an iterative method that models the effects of illumination changes over time is proposed to improve the accuracy of BTFs for re-identification. In [177] the re-identification problem is posed as a classification problem in the feature space formed of concatenated features of persons viewed in two different cameras.

- Recent Trends: Some recent methods have addressed a major shortcoming of many of the above approaches that they rely on the appearance features of the person, and thus are not robust to simple issues like change of clothing. For example, the usage of soft biometric cues as features, alternative or complementary attributes to classical appearance-based re-id, is envisaged as one of the current trends in this field. Actually, soft biometrics systems mostly deal with subjects that do not have a strong collaborative behavior, eventually more suitable for more realistic scenarios. In such cases, discriminant cues can be extracted from range data acquired by RGB-D cameras, such as the MS Kinect or Asus Xtion PRO, which are able to acquire depth information in a fast and affordable way. The idea here is to consider more implicit human body characteristics, in particular, to extract a set of features computed directly on the range measurements given by the sensor related to specific anthropometric measurements [178].

  Another interesting research line regards the use of a pan-tilt-zoom (PTZ) camera for re-id [179]. The idea here is to build a person descriptor which contains both typical information (full body, specific areas) at a certain resolution, and also specific peculiar human regions at higher resolution (e.g., a leg or an arm), to capture discriminant characteristics (e.g., a tattoo) able to uniquely distinguish among the subjects. This of course entails another difficulty of efficiently maneuvering the PTZ camera for person detection and tracking. A summary of the main research contributions in this field is given below in Tables 1 and 2.

- Evaluation Methods: Re-id algorithms are typically tested considering ad hoc datasets consisting of a fixed number of individuals, each one represented by several (minimum two) image windows or bounding boxes surrounding the person, typically in different poses (e.g., frontal, back, profile, etc.). The actual re-identification is performed by considering one instance of an individual, the so-called "probe", which is compared and matched against all the other instances of all subjects in the dataset, the so-called gallery set. Hence, given a set of probes, a re-id method provides a ranking of a certain number of subjects in the gallery set, typically ranging in the upper tens, depending on the used dataset. Algorithms performance is commonly calculated by the

| **Authors** | **Year** | **Approach** | **Features** | **Temporal Information** | **Representation** |
|---|---|---|---|---|---|
| Javed et al. [180] | 2005 | Feature Transformation | Color | Used | Color appearance with Color brightness transfer function (BTF) |
| Gilbert et al. [173] | 2006 | Feature Transformation | Color | Used | Consensus-color conversion of munsell color space with color transformation matrix |
| Gheissari et al. [181] | 2006 | Discriminative Signature | Color and shape | Used | Graph partition based representation |
| Hu et al. [182] | 2006 | Discriminative Signature | Geometry | Used | Principal axis with segmentation |
| Wang et al. [183] | 2007 | Discriminative Signature | Color, gradients and shape | Not used | Co-occurrence spatial context |
| Chen et al. [184] | 2008 | Feature Transformation | Color | Used | Color appearance with temporal color brightness transform and spatial information |
| Prosser et al. [185] | 2008 | Feature Transformation | Color | Used | Color appearance with temporal color brightness transform and spatial information |
| Javed et al. [172] | 2008 | Feature Transformation | Color | Used | Color appearance with spatial temporal color brightness transform and spatial information |
| Gray and Tao [186] | 2008 | Discriminative Signature | Color, gradients and filters | Not used | Selected histogram features by Adaboost |
| Zheng et al. [187] | 2009 | Discriminative Signature | Color and gradients | Not used | Grouping as dynamic spatial context |
| Bak et al. [188] | 2010 | Discriminative Signature | Color | Not used | Covariance matrix between parts |
| Farenzena et al. [189] | 2010 | Discriminative Signature | Color and structure | Not used | Symmetry-based ensemble of local features |

Table 1: Main Contributions in the Fields of Person re-identification.

| Authors | Year | Approach | Features | Temporal Information | Representation |
|---|---|---|---|---|---|
| Prosser et al. [190] | 2010 | Metric Learning | Color, gradient, filters | Not used | Quantified histogram feature by RankSVM |
| Cheng et al. [147] | 2011 | Discriminative Signature | Color and structure | Not used | Pictorial structures modeling |
| Dikmen et al. [158] | 2011 | Metric Learning | Color | Not used | Large Margin Nearest Neighbor with Rejection on densely sampled color histogram features |
| Kostinger et al. [160] | 2012 | Metric Learning | Color and texture filters | Not used | KISS Metric Learning on densely sampled color and texture features |
| Datta et al. [175] | 2012 | Feature Transformation | Color and shape | Not used | Weighted brightness transfer function on color and shape histogram features |
| Gala and Shah [153] | 2012 | Discriminative Signature | Color and structure | Used | Active color model with representative meta colors extracted over multiple frames |
| Kviatkovsky et al. [155] | 2013 | Discriminative Signature | Color | Not used | Capturing color shape distribution in the log-chromaticity color space using shape context |
| Zhao et al. [156] | 2013 | Discriminative Signature | Color and gradients | Not used | Bi-directional weighted matching on densely sampled color and SIFT features |
| Li et al. [191] | 2013 | Feature Transformation | Color, shape and texture filters | Not used | Metric learning on locally aligned color, shape and texture histogram features |
| Zheng et al. [192] | 2013 | Metric Learning | Color and texture filters | Not used | Relative distance comparison on color and texture features |

Table 2: Main Contributions in the Fields of Person re-identification.

recognition rate estimated by the Cumulative Matching Characteristic (CMC) curve, and the normalized Area Under Curve (nAUC) score for the CMC curve. The CMC curve is a plot of the recognition performance vs. the ranking score and represents the expectation of finding the correct match in the top n matches; nAUC gives an overall score of how well a re-identification method performs overall.

The most used image datasets are VIPER, iLIDS, ETHZ, and CAVIAR4REID, which differ for a number of characteristics: number of subjects considered (in the order of hundreds), number of instances per subject (from 2 to 10), pose and lighting variations, severity of the occlusions, number of viewpoints (cameras), window resolution (from $17 \times 39$ up to $72 \times 144$, $48 \times 128$, $64 \times 128$, etc.). Depending on the number of instances of a subject in the dataset, the re-id procedure can be performed in single-shot or multiple-shot modality. In the former case, we have only 2 instances, one in the probe and the other in the gallery set, so person signatures are calculated and matched only considering these 2 image windows. In the latter, more than two instances per subject are given, and the process may take into account this richer information (e.g., using more instances for calculating the descriptor or to make the extracted features more statistically significant). Also, 3D datasets are now becoming available for re-id research. A summary of the characteristics of the major datasets is given in Table 3 and detailed in a recent review paper [145].

| **Dataset** | **People** | **Image Info** | **Cameras** | **Additional Info** |
|---|---|---|---|---|
| ETHZ (Seq #1, Seq #2, Seq #3) | (83, 35, 28) | Images: (4856, 1690, 1762) Avg. images per person per camera: (59, 48, 63) | (2, 2, 2) | Scenario: outdoor Challenges: color changes, occlusions, sptial resolution http://homepages.dcc.ufmg.br/∼william |
| CAVIAR4REID | 72 (50) | 1220 (1000) Avg. images per person per camera: 10 (10) | 2 | Scenario: indoor Challenges: viewpoint variation, color changes, spatial resolution www.lorisbazzani.info |
| WARD | 70 | Images: 4786 Avg. images per person per camera: 69 | 3 | Scenario: outdoor Challenges: viewpoint variations, spatial resolution, color changes http://users.dimi.uniud.it/∼niki.martinel |
| VIPeR | 632 | Images: 1264 Avg. images per person per camera: 1 | 2 | Scenario: outdoor Challenges: viewpoint variation, color changes http://vision.soe.ucsc.edu/node/17 |
| SAIVT | 152 | Images: 64778 Avg. images per person per camera: 53 | 8 | Scenario: building environment Challenges: viewpoint variation, illumination variation https://wiki.qut.edu.au/display/saivt/SAIVT-SoftBio+Database |
| I-LIDS (MCTS) | 119 | Images: 476 Avg. images per person per camera: 4 | 5 | Scenario: Outdoor transport hub Challenges: viewpoint variation, Clutter www.ilids.co.uk |

Table 3: A summary of the major datasets used in re-identification experimentation.

## 6.3 Solved Problems

The typical input data consists of high quality images without significant occlusion and clutter, and the scenarios considered are rather constrained and cooperative. Constrained could mean that the people are limited to certain regions (e.g., walkways) and their features can be identified and extracted. Cooperative means that they are not trying to evade identification; in the extreme, this could mean that they are willing to provide their images in a fixed setup (e.g., a security checkpoint).

In such conditions, the pre-processing part of the re-id process is relatively reliable, that is, detection and tracking moving objects, can be considered to be solved. Accurate foreground and body part extraction can be considered as effective; hence feature extraction, the building of the signature descriptor, and the consequent actual recognition, are problems which are going to be solved in the short-term.

Better (i.e., more realistic) experimental datasets are actually needed to test and validate the current algorithms in real-world applications. One aspect that should be considered is data collected over multiple cameras and longer time periods. That will help analyze how consistent the results are even in these relatively constrained environments. An example of the kinds of images on which reliable results are possible in the short-term is given below in Figure 26.

Figure 26: Examples of images on which reliable re-identification results can be obtained in the short term (ViPER dataset).

## 6.4 Near-term Solvable Problems

If we consider input images of slightly lower quality, with some partial occlusions, still in a cooperative but relatively unconstrained scenarios, other problems arise which require a more thorough and longer-term investigation. Examples could be people walking on a not-too-busy street and not objecting to having their images taken.

The extraction of reliable features is a major challenge when dealing with data with partial occlusions and large variations of appearance. In such scenarios, moderate body pose

variations and the associated appearance and resolution changes may require the use of more robust cues, like 3D features and soft biometric data, including distinctive signs which may characterize univocally an individual (e.g., a tattoo). The integration of attributes (e.g., a carrying bag), and in general, of contextual information will actually become necessary for disambiguating individuals in a more robust and effective way. Given the more realistic conditions, detection and tracking cannot be considered anymore to be reliable procedures, so the propagation of errors from such modules should be taken into account.

The necessity of better benchmark, more practical, datasets and related performance measures, still remains a requirement for the deployment of the re-id technology in the real world. In these cases, one may consider the novelty detection problem, i.e., to decide when an individual (probe) is not in the gallery and the automatic enrollment of a probe image in the gallery set. Besides, one may consider inclusion of human in the loop and this will entail additional issues to be studied like, for instance, the human feedback for the refinement of the query. Active image acquisition, possibly through the human, is another aspect that can be studied.

Currently, there are no good datasets to rigorously study these problems. A task will be to collect such datasets, annotate them, and make them available widely. The design of such datasets needs to be carefully thought, if they are acquired in staged scenarios. Examples of some relatively challenging scenarios are given below in Figure 27. Current methods are unable to recognize the people in the top and bottom rows to be the same, due to large changes in appearance (first three columns), or partial occlusion and clutter in the last two.

Figure 27: Examples of some images that current methods have difficulty in reidentifying.

## 6.5 Problems that Need Long-term Investments

Full real-world scenarios, low-quality images, unconstrained and uncooperative conditions are challenges that will need a longer time horizon. This will involve dealing with natural videos with high clutter in the data, and severe variations in the environmental conditions.

An example could be a busy city scene, where a person needs to be recognized as he walks through several blocks with large blind areas in between.

The main task identified above robust feature extraction remains the key and will need to be achieved in far more challenging conditions. This may call for the development of novel features. Tools that could be useful for this purpose include image restoration (e.g., super resolution) techniques to improve the quality of the acquired images/videos. It is to be expected that a larger use of context, like the joint re-id of groups and individuals, can be helpful. Semantically meaningful attributes could play a role in providing the required robustness. The development of online learning can be a major step in the feature extraction problem. Methods (like deep learning and sparse coding) which can automatically learn the best features, rather than using hand-engineered ones, hold promise in this respect. The use of multi-modal multi-sensory input (beyond optical, e.g., multi-spectral, infrared) should also be considered to cope with real scenarios, as well as the potential exploitation of active acquisition systems and mobile platforms for collecting more information-rich data. This could allow capturing salient parts of the human body as a way to acquire features that would enable recognizing people in such harsh scenarios. All these techniques should scale gracefully with large numbers of cameras, and cope with wider space-time horizons.

The construction of large datasets "in the wild" will be a necessity to validate the developed re-id technologies. However, collecting such datasets that will be meaningful and annotating them reliably will be a challenge. An example of a truly challenging scenario for long-term research is shown in Figure 28 below.

Figure 28: A truly challenging scenario for re-identification which could be considered as a long-term research issue.

# 7 Human Activity Understanding (Detection and Recognition) in Video[5]

## 7.1 Introduction

Events, actions, interactions, activities, and behaviors have been used interchangeably in the literature, and there is no agreement on the precise definition of each term. In a recent paper

[5]Anthony Hoogs, Ram Nevatia, Mubarak Shah and Rene Vidal

[193] an attempt was made to provide a hierarchical model for complex event recognition shown in Figure 29. In this part of the report, we present the following definitions and descriptions directly from [193]. See [194, 195, 196, 197, 198, 199, 200, 201] for more details on various recent activity and action recognition methods.

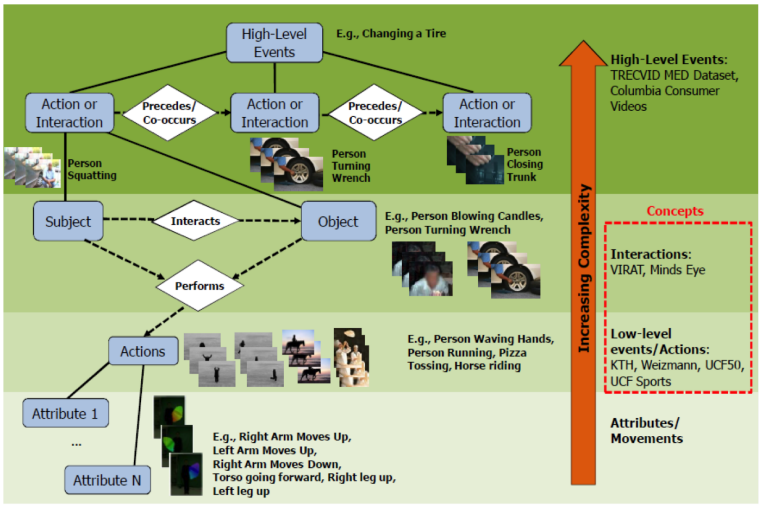



Figure 29: A taxonomy of semantic categories in videos, with increased complexity from bottom to top. Attributes are basic components (e.g. movements) of actions, while actions are key elements of interactions. High-level events lie on top of the hierarchy, which contain (normally multiple) complex actions and interactions evolving over time.

Movement is the lowest level description: "an entity (e.g. hand) is moved with large displacement in right direction with slow speed". Movements can also be referred to as attributes which have been recently used in human action recognition following their successful use in face recognition in a single image. Next are activities or actions, which are sequences of movements (e.g. "hand moving to right followed by hand moving to left", which is a "waving" action). An action has a more meaningful interpretation and is often performed by entities (e.g., human, animal, and vehicle). An action can also be performed between two or more entities, which is commonly referred to as an interaction (e.g. "person lifts an object", "person kisses another person", car enters facility, etc.). Motion verbs can also be used to describe interactions. Recently the Mind's eye dataset was released under a DARPA program which contains many motion verbs such as "approach", "lift" etc. In this hierarchy, concepts span across both actions and interactions. In general, concept is a loaded word, which has been used to represent objects, scenes, and events, such as those defined in LSCOM (Large-Scale Concept Ontology for Multimedia). Finally, at the top level of the hierarchy, we have complex or high-level events that have larger temporal durations and consist of a sequence of interactions or stand-alone actions, e.g., an event "changing a vehicle tire" contains a sequence of interactions such as "person opening a trunk" and "person using a wrench", followed by actions such as "squatting" and so on. Similarly, another complex event such as "birthday party" may involve actions like "person clapping" and "person singing", followed by interactions like "person blowing candles" and "person cutting a cake".

Note that although we have attempted to encapsulate most semantic components of complex events in a single hierarchy, because of the polysemous nature of the words, adopting the same terminologies in the research community is an impossible objective to achieve. Having said that, we set the context of event recognition as the detection of temporal and spatial locations of the complex event in the video sequence. In a simplified case, e.g., when the temporal segmentation of a video sequence into clips has been achieved, or each video contains only one event and precise spatial localization is not important, the even recognition problem stated above reduces to a video classification problem.

Besides the above hierarchy, there is another class of events that are global in nature and involve multiple actors (or agents). For instance, crowd behaviors involving a large number of people are global events. There are at least five different events can be identified in crowds: Bottlenecksmany pedestrians/vehicles from various points in the scene enter through one narrow passage; Fountainheads many pedestrians/vehicles emerge from a narrow passage only to separate in many directions; Lanes-many pedestrians/vehicles moving at the same speeds in the same direction; Arches or Ringscollective motion is curved or circular and Blockingdesired movement of groups of pedestrians is somehow prohibited. In traffic scene global events involving multiple vehicles include: Straight, Left turn, Right turn, Acceleration, Deceleration, Right U turn, Left U turn, Circular, Divergence, Convergence, Stop and start. Another interesting domain for multi-agent events is the American football. Plays in American football involve very complex movement and interactions of players. These plays have been well defined by NCAA, the challenge is to be automatically detect such plays in videos. The current reported work in the literature is able to classify seven plays: Left Run, Middle Run, Right Run, Roll Out Pass, Short Pass, Deep Pass and Left Pass.

Given this general introduction and definitions, in the next three sub-sections we discuss the problems in the context of human activity understanding which are considered to be solved (Short Term); the problems which will be solved in next three years (Mid-Term) and the finally problems which are unsolved (Long Term).

## 7.2 Solved Problems

A simplified formulation of the action classification problem can be stated as follows: Given a trimmed video containing a particular action from a known set of action classes, automatically determine what class the video belongs to. We consider that this simplified action classification problem in somewhat solved. This kind of analysis does not provide any spatial or temporal localization of an action. Most approaches use global methods like bag of words and employ descriptors around local spatiotemporal interest points and dense trajectories. The classification is performed by training Support Vector Machine classifiers in a supervised manner.

In a recent Action Classification contest, THUMOS [202], held at ICCV 2013, top ranked methods were able to perform classification with accuracy of more than 85% on a data set covering 101 (UCF-101) different action classes. The UCF-101 data set contains roughly three hours of video involving 13,320 short video clips ranging in length from 7 seconds to 70 seconds recorded at 25 frames per second with resolution of 320 by 240. Action classes are divided into five main groups consisting of Human-Object actions, Body Motion actions, Playing Instrument actions, Human-Human actions and Single Person Sport action. The

complete list is given in [202].

Current approaches to action classification in context of bag of words involve three main steps: Feature Extraction, Pooling and Encoding, and Classification. The main progress during the last few years has been in developing of robust features, which can be reliably extracted from videos and used in action classification. These features include: STIP (Spatiotemporal Interest Point detector and descriptor), MBH (Motion Boundary Histogram), HOG (Histogram of Oriented Gradient), HOF (Histogram of Optical Flow), DTF (Dense Trajectory Features) etc. Recently, there has been lots of interest in data driven features using deep learning methods. The idea is that instead of using hand crafted features like MBH, the features are learned from the data. The deep learning approach has been very successful in discovering features and using them in classifiers for the ImageNet challenge, which involves images of more than 1,000 classes. However, the impact of deep learning approaches on video has been a bit limited due to the lack of a large number of annotated videos.

For pooling and encoding, the Fisher Vector based approach has been very successful compared to just standard k-means clustering typically used in bag of words. However, for classification purpose almost all approaches use SVM.

## 7.3 Near-term Solvable Problems

Current action data sets, including UCF-101, have one weakness in common: they provide only temporally segmented videos. Therefore, the action recognition methods evaluated on such benchmarks suffer from the restrictive assumption that the action of interest must have been temporally localized in advance. This limitation is identified to be a major barrier in developing practical action recognition techniques in numerous research publications. As the result of the absence of a pragmatic large dataset, very little amount of effort towards performing action localization and recognition in realistic scenarios has been reported to date. Therefore, action localization (spatial and temporal) in realistic videos containing large number of classes with decent accuracy is still an unsolved problem. However, we expect this problem will be solved in the next three years.

There are some promising approaches being pursued in this direction. In particular, we expect that good progress will be made in the next three years in recognizing complex actions comprising a sequence of simple actions. Several approaches to this task are being investigated. Most traditional is to model the complex action by a dynamical model such as a Hidden Markov Model (HMM) or a Switched Linear Dynamical System (SLDS), where each state belongs to a primitive action. Finding the most likely path in such a model provides the most likely sequence of actions. In this way, a video is segmented into a sequence of actions while at the same time each action is classified. One limitation of these approaches is that temporal relationships are limited by the Markov assumption of HMMs and SLDSs. More recent work uses Markov and Semi-Markov Conditional Random Field (MsM-CRF) models, which are able to perform joint action segmentation and classification by capturing long range temporal relationships among actions. However, even though these models contain a representation of uncertainty, they can be overwhelmed by missing or completely erroneous data. A more robust technique is use of Conditional Bayesian Networks (CBNs) that model only the presence of critical states in a temporal order but rather loose restrictions on their

transitions. These methods have been applied to surveillance videos where the imaging conditions and action sequences are relatively invariant.

The aforementioned methods provide temporal localization of an action, but not spatial localization. One approach to both spatial and temporal localization is the Spatiotemporal Deformable Part Model (SDPM), which is an extension of the popular DPM model for single object detection. In DPM paradigm sliding window is used to search through to localize and detect an object. In SDPM, the actions are treated as spatiotemporal patterns and a deformable model with discriminative parts, which are in fact discriminative 3D sub-volumes, is learnt for each action class from the training data. Similar to original DPM model SDMP run across the video spatially and temporally in a sliding window fashion to detect and localize an action. This approach is shown to perform well in spatial localization of action and cope with the limited background clutter on simple datasets like UCF Sports and MSR.

Another approach that is gaining interest is video segmentation followed by action classification. In this class of approaches, super voxel segmentation of a video clip is first performed. An action is then represented as a group of super voxels and some features are computed for each super voxel and are used to train a classifier. However, this approach heavily relies on the quality of video segmentation. In fact segmentation (image and video) continues to remain a fundamental problem in computer vision, and efficient and robust solution to this problem will help solve many other problems in computer vision. In particular, video segmentation and tracking are related, i.e., tracking is a special case of in video segmentation. In video segmentation, essentially each region including background and foreground are tracked throughout the video frames instead of a few object regions in tracking.

Another limitation of most action recognition methods is that actions are pretty simple, e.g., golf swing, drinking or smoking. As we start to consider videos in the wild, the variations become more extreme. For such videos, most robust methods have proven to be those using statistics of local features, much as for the primitive actions. However, methods have started to emerge that learn discriminative segments from a collection of highly varying videos and use these segments to classify the high level activities. In current stage, these methods can improve the performance of lower level classifiers by fusion. We expect that in the next couple of years, they will begin to significantly outperform the lower level methods. The discriminative segments, however, do not necessarily correspond to semantic entities so mapping them to natural descriptions can be a challenge.

Another limitation in current approaches is the requirement of large number of training examples. Since almost all methods use classifiers like Support Vector Machine, which require labelled positive and negative examples for each class. This involves manual annotation of a large number of sample videos, which is a pretty challenging task if the number of classes is large. For action localization, task annotation at the bounding box level for each actor in each frame is required, not just labeling of a video as a positive or negative example for a particular class. Due to this manual effort, there does not exist such annotated data with large number of classes. For a long time, the only reasonable sized annotated data set for action localization purposes has been UCF Sports containing 9 actions. In the THUMOS13 contest, annotations for an action data set containing 24 actions (UCF sports is a sub-set of this data set) were provided. However, due to the complexity of the action localization task, there was no single submission in this track.

In order to deal with the above limitation of large number of training examples, there is some interesting work being pursued under the title of k-short learning. These methods emphasize the rich representation to be used in action classification, which is able to distinguish different classes by using only a few examples. We expect this line of research will continue in next few years and to result in methods requiring a few training examples.

A related problem is cross data-set generalization of current methods. Since the classifiers are trained on specific data sets, they do not generalize to some other data set, e.g. training on UCF-50 and testing on HMDB-51. The performance in that case is much lower. This difference in performance is even noticeable when the training is done on static images, e.g., training on the Imagenet data set and testing on frames of videos. Compute Vision researchers have realized this issue and currently several interesting approaches are being pursued in this direction, and we hope a good progress will made in next three to five years. One approach is domain adaptation, where the representations learned from the training set are adapted to the test set by transforming both data to some common latent space. Another approach, which has not been explored yet, is that classification techniques need to revisit their basic assumption that the training examples are identically and independently distributed.

## 7.4 Problems that Need Long-term Investments

The ultimate goal of activity and action understanding in Computer Vision is to be able to provide explanations and descriptions of an action or event captured in a video. This kind of analysis is very important for end users, e.g., video analysts. As an example, take a TRECVID event such as "Basket Ball Going through the Hoop. In order to really provide an explanation of this video, the system should be able to detect a "basketball and the hoop and be able to track the "ball and detect if the "basketball did go through the "hoop. In early years, the approaches essentially identified a key frame of a video shot, extracted some global features like color histogram and trained an SVM classifier. More recent approaches use motion information, e.g., descriptors around spatiotemporal interest points or dense trajectory features like MBH. However, still the approaches being used are global in nature, since global histogram for whole clip following bag of words paradigm is used to train a classifier. These methods do not have any notion of "ball, "basket or the event when the ball goes through the hoop. The problem is that an SVM is a black box, which is only able to separate feature representations of positive and negative examples. However, mapping of a video to a feature vector like a bag of words loses all its semantic meaning since it is just a collection of numbers.

To understand and explain a video, it is important to have a rich representation of each event, action or activity in terms of objects, actions and scenes, which can be used to describe an event in natural language. That kind of description will be very useful for video analysts. There is a recent task: Multimedia Event Recounting (MER) under NIST TRECVID, which is a very good starting point for addressing this problem. For high quality descriptions, ideally, we should have robust estimation of human actors positions and poses and an understanding of the objects in the environment, particularly the ones that the humans interact with. While this task is challenging for any domain, it is particularly so for the videos in the wild. The first difficulty is that the choice of vocabulary, i.e. the ontology

of the events itself needs to be defined. Then, there is a large variety of objects that can be present: consider actions talking about "appliances or "animals and the many varieties that may be present. Finally, the performance of specific object and action detectors is still rather low.

There are emerging efforts that aim to bridge the gap between the semantics required by the high level description and what can be extracted from the lower level detectors. Essentially, the task becomes a data-driven learning of the high level concepts from a distribution of the responses of the lower level detectors. The outputs of the lower level detectors can be viewed as a mapping of the original video signal to a much smaller dimensional set of meaningful entities. Note that MER can and should use information from multiple modalities that include, besides the video content, concepts inferred from text and audio, as the latter map much more directly to linguistic entities. Thus, we can anticipate steady improvements in MER performance, along with improvements in detection accuracy.

As discussed before almost all current approaches to activity and action classification are supervised, i.e., these approaches require some kind of training data. Computer Vision research started with model-based and knowledge based approaches. In these approaches training data is not required. For instance, in Lowes model based approach for object recognition, in to order recognize a car 3D model of the car is projected to the image plane and projected model straight lines are matched with the straight lines in the image and pose is estimated. Different views of car can be easily generated given a 3D model, therefore no training data is required. However, these model based approaches did not work that well, because methods for low level features (like detection of straight lines) were not that robust. Now that great progress has been made in robust extraction of low level features it will be interesting to revive model and knowledge based approaches in computer vision. Moreover, to extend 3D model based approaches to multiple categories, we envision learning such 3D models of a category from multiple 2D images.

In the context of action recognition, from human knowledge the "kicking action can be described as a sequence of left arm moves up, left leg extends forward, left leg extends back, left leg retracts and right leg retracts. This kind of representation does not require any training and beside action classification it can also be used for explanation. The long term challenge for computer vision researchers is to develop approaches for human activity and action understanding which do not require any training and which can generalize to diverse datasets and provide explanation and recounting of actions and activities in a video.

# 8 Semantic Summarization (Attribute-based Scene Tagging[6]

Image-based semantic scene summarization is the process of assigning attributes to a viewed scene in order to provide an expressive meaning of the scene content. The focus of this topic is on spatial scene interpretation properties i.e., "what does this scene contain?", "what are the characteristics of this scene and objects within it?" A separate topic focuses on activity understanding based on the motion and interactions between entities in a scene,

[6]Gil Ettinger, Jianbo Shi, Martial Hebert

as extracted from video data. The semantic summarization process may extract information from either images (photographs) or video, but any video processing focuses primarily on multi-frame spatial scene understanding rather than motion characterization. For semantic summarization, an event summary may consist of a group of $> 10$ people standing on the sidewalk of an urban setting. On the other hand, activity understanding (which is not the focus of this semantic summarization topic) is focused on the motion attribution of actors in the scene, such as "a protest by a growing group of people who moved from the SE corner of the plaza to the NW corner, where policemen were standing, sparking a violent clash."

Semantic summarization includes the following types of attributes:

- Object Attributes: Properties of detected objects in the scene, such as types, counts, sizes, poses and appearance attributes. Common objects of interest are people and vehicles. Object types may be given at coarse or fine levels of specificity e.g., animal >> dog >> golden retriever.

- Region Attributes: Segmentation and classification of scene regions, in a 2D image projection or in a 3D reconstruction of the scene. Region classification include such labels as: ground, sky, vertical surface, building, waterway, road, field, forest, Region attributes may also include functional labels such as: entrance, on-ramp, intersection,

- Relational Attributes: Spatial relationships between entities in the scene, consisting of object-to-object and object-to-scene relationships. These attributes include such relationships as: "person standing on top of car", "person sitting on motorcycle", "person crouching under table", "person standing in creek"

- Location Attributes: Specific geolocation or more general location attributes, such as region of the world or type of scene (desert, jungle, beach,).

- Temporal Attributes: Absolute time when image was taken or relative time indications, such as time-of-day, day-of-week, season.

Semantic summarization is a complex reasoning process that faces the following challenges:

- Developing the terminology of "semantics": Semantic summarization is used to communicate scene content with people by using words and phrases, but different people may use different words to describe the same, or similar concepts. As a result, comprehensive ontologies, such as WordNet and ImageNet, have been developed to express relationships between words and image exemplars of those words. We do not focus on the ontology of semantics for this FIVA application.

- Extracting 3D structure from 2D image(s): The application of physical constraints is powerful technique for generating and refining feasible semantic scene interpretations. The estimation of relative or absolute 3D structural relationships between hypothesized scene objects and between objects and the scene provide rich information for assessing metric sizes of objects, assessing height of objects above the ground, evaluating visibility/occlusion effects, and verifying shadowing effects.

- Characterizing scene entities by appearance, function and relationship to other entities in the scene: While confidently detecting objects in a scene is a primary goal of semantic summarization, it is generally desirable to also estimate object attributes that characterize (1) appearance, such as color and pose, (2) function, such as road (i.e., support vehicle traffic), sidewalk (i.e., supporting pedestrian traffic), obstacle, pathway, and (3) spatial relationships between objects, such as "person holding tool", "person riding motorcycle", "person on top of car". These attributes not only provide meaningful semantics, but also improve a systems ability to detect the described objects since these attributes constrain the visual appearance of the objects in the imagery.

- Detecting objects that are partially obscured, in shadow, in dense groups: Realistic scene understanding problems are often characterized by partially visible signatures of objects of interest, where the visibility of the full objects may be partially blocked by scene entities or other objects or may be rendered imperceptible by poor lighting or shadows. Groups of objects, such as crowds of people, may have salient signatures that make detection of the group possible even through the individual components may not be detectable on their own. While humans may be able to indirectly infer the presence of objects that are totally occluded in a scene, the goal of semantic summarization is to only detect and identify objects for which there is some direct evidence in the imagery. We would like to withstand as much partial occlusion as possible in detecting objects in a scene, and leave the problem of inferring non-visible object presence to a separate challenge for higher level reasoning components.

- Exploiting external (non visual) knowledge sources: The semantic summarization task can be aided by data that may be available to constrain the scale and breadth of the content search problem. Two key sources of external data that may be available are: (1) metadata collected by the camera, such as geo-location, time, camera calibration parameters, and (2) reference data, such as terrain maps, elevation maps, land use maps, priors on object type populations. Such information may be of significant utility, but can also be inaccurate and/or out of date. Semantic summarization techniques should leverage such external knowledge sources if available, but not depend on the availability, nor accuracy, of any one of them.

- Multi-scale semantic scene understanding to address variable effective image resolutions and varying levels of detail: The scale of objects in an image, in terms of pixels-on-target and occupied fraction of the image, can significantly vary. Even in a single image, multiple instances of a single object type may exhibit extensive depth diversity. Image quality may also impact effective resolution. As a result, semantic summarization capabilities must address the highly variable levels of detail that are available and exhibit graceful performance degradation with lower effective resolution.

- Detecting anomalies in the scene: Most scene understanding algorithms are generally targeted to work for the common situations that is, find the common appearances, in the common poses, in the common locations, based on statistical priors and learned models training on exemplar data. But it may be the unusual object instance that may actually be of most interest in some applications. That is, the person on the rooftop,

the awkward pose of a person holding a heavy bag, the presence of a vehicle in an unusual location may be anomalies of interest that are difficult to identify precisely because they are unusual.

- Estimating confidence of scene interpretations: A meaningful confidence estimate should be associated with resultant semantic summarizations in order to make the results most useful. Computing posterior beliefs on scene interpretations is a challenging problem since underlying knowledge bases are incomplete and independence assumptions that decompose belief estimation to tractable calculations are not always valid.

## 8.1 Solved Problems

Robust solutions to the semantic summarization problem, which provide automated high-performance solutions across a wide range of operating conditions, do not yet exist. Recent progress in computer vision research, though, has developed promising foundational building blocks for generating semantic scene summarizations. These include general purpose object detection algorithms as well as image segmentation and labeling algorithms, as described in the following sub-sections. The basic semantic summarization capabilities enabled by these algorithms today are:

- Detection and classification of unoccluded foreground objects in ground-based images with at least 1000 pixels on target.
- Semantic image region labeling of coarse geometric scene layout (e.g., ground, sky, buildings), particularly when applied to video sequences rather than individual images.

These basic semantic capabilities may not provide comprehensive semantic scene descriptions, but even as incomplete and/or uncertain descriptions they enable users to develop application-specific image analysis products and triage large volumes of images of interest much more efficiently than could be done manually.

### 8.1.1 Unoccluded Object Detection

Machine learning techniques have been adapted and applied to the problem of detecting large classes of entities in 2D images, primarily collected with ground-based cameras. Object detection is a vital enabler for semantic summarization applications. Up until recently the state-of-the-art has been the Deformable Part Model (DPM) detector [203], which applies an SVM classifier to assess HOG features in a hierarchical manner. The software for DPM is available at: http://www.cs.berkeley.edu/ rbg/latent/voc-release5.tgz. A key benefit of this software is that the training software is available, as well as the online detection software. But he most successful recent object detection work is the Regions with Convolutional Neural Nets (R-CNN) method [204], which integrates CNN features with SVM classification. After several years of no community improvement on object detection benchmark tests, such as PASCAL VOC, the 2014 R-CNN results showed a mean average precision (mAP) improvement of more than 30%, achieving a mAP of 53.3% across a broad range of object

classes. Detection performance covers a wide range of objects, including people, many vehicle classes, many animal classes, and many household objects. Software for R-CNN is available at: http://www.cs.berkeley.edu/∼rbg/rcnn. While the R-CNN object detection performance is significantly better than previous systems, it is insufficient to achieve effective results on its own for practical image analysis problems. That is, many true detections are missed at manageable false alarm rates. The types of errors made by R-CNN, as reported in [204], in order of occurrence (most to least) are: poor localization, confusion with similar object category, confusion with dissimilar object category, and false positive on background. Additionally, detecting partially occluded objects remains a challenging problem, with detection performance degrading quickly as occlusion rate increases. A final issue is the effective resolution of imagery on objects of interest. Sufficient pixels-on-target, generally ¿ 1000 pixels, is generally needed to reliably detect objects. But since detection techniques are relatively robust to scale and the size of camera focal plane arrays is rapidly increasing, image resolution is becoming less of an issue. Sample object detections results for DPM and R-CNN are shown in Figure 30.

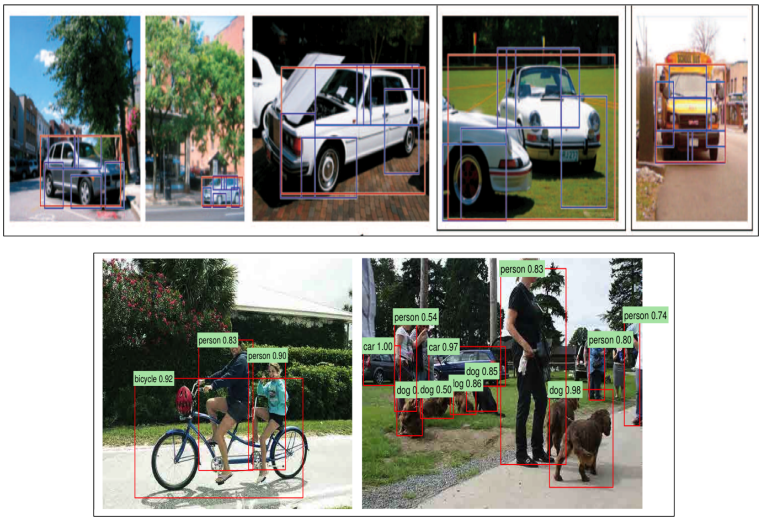



Figure 30: Sample results for object detection algorithms: DPM top from [3]; R-CNN bottom from [204].

### 8.1.2 Coarse Image Segmentation and Labeling

Meaningful progress has been made in addressing the problem of segmenting and labeling coherent regions in images. A robust technique for labeling ground, sky and vertical surfaces in ground-based imagery is the Geometric Context algorithm [205]. This algorithm applies machine learning techniques to a collection of color, texture, location/shape, and geometry features to generate meaningful scene geometry labels for groupings of image superpixels. The Geometric Context software is available at:
http://web.engr.illinois.edu/∼dhoiem/projects/software.html. Given the relatively limited label vocabulary of the Geometric Context results, it performs reliably across a wide range

of operating conditions. Many other techniques attempt to label image regions with a much finer set of region classes. Promising recent image labeling algorithms include the multiscale scene parting algorithm [206] and the Superparsing technique [207]. The latter performs particularly well on large scale regions such as: building, tree, sky, road, sidewalk. Software for the Superparsing technique is available at: http://www.cs.unc.edu/~jtighe/Papers/ECCV10/eccv10-jtighe-code.zip. Since labeling of single images is in general relatively noisy, the segmentation techniques may be applied to video sequences with the application of temporal coherence to converge on the likeliest segmentation and produce more reliable results. Sample image segmentation and labeling results are shown in Figure 31.


Figure 31: Sample image segmentation and labeling results [205, 207].

## 8.2 Near-term Solvable Problems

In this section, we summarize semantic summarization capabilities for which ongoing research efforts are developing emerging solutions. Additional limited investment in productizing and engineering practical and robust systems that incorporate these capabilities should yield deployable solutions in a 1-3 year timeframe. These capabilities include:

- Detailed Image Region Labeling: Image segmentation and labeling (image parsing) continues to be an active area of computer vision research. Successful techniques are focusing on incorporating contextual information to resolve segmentation and labeling ambiguities. In essence, these algorithms are moving towards developing physically consistent scene interpretations since the 2D (super)pixel-level segmentation does not work effectively for identifying relatively small regions and resolving region labels in cluttered environments.

  One promising technique to provide reliable and robust image region labeling is the fusion of 2D image segmentation/labeling algorithms with object detection algorithms [208], as shown in Figure 32. By combining results from trained object detectors with image segmentation/labeling results, Tighe and Lazebnik are able to generate regions

segmentations that are more accurate as well as better labeled than earlier techniques. Their software is available at: http://www.cs.unc.edu/SuperParsing.

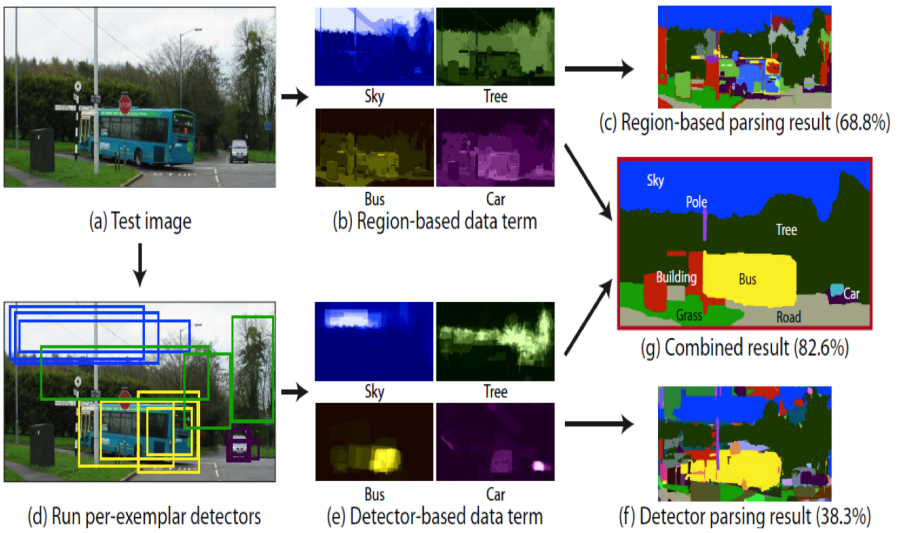



Figure 32: Joint image parsing with regions and per-exemplar detectors [208] shows promising results for reliable and robust image segmentation and labeling.

Another promising image segmentation and labeling approach, which is aimed at identifying relatively small image regions, such as downrange people and vehicles, is the context-driven scene parsing algorithm [209]. This technique reportedly improves the segmentation of the relatively small regions by over 10% over conventional approaches. A sample result of this algorithm is shown in Figure 33.

- Semantic 3D Reconstruction: 3D scene geometry is a key component of semantic summarization on its own, but is also an enabler for improved image segmentation, object detection, false alarm suppression and inter-object relationship discovery. Absolute or relative 3D scene modeling, particularly if it can be done from single images, enables metric size analysis, visibility analysis, shadow/occlusion confirmation, and inter-object relationship estimation. Desirable scene relationships such as inter-object proximity require 3D reasoning to determine that even though objects may appear near other in the 2D image plane, they may be far apart from each other in the 3D world.

  An early technique for estimating 3D depth from a single image is Make3D [210]. This method applies a trained Markov Random Field to infer a set of plane parameters that capture both the 3D location and 3D orientation of image patches  sample results are shown in Figure 34. Make3D software is available at: http://make3d.cs.cornell.edu/code.html. More robust, but qualitative, 3D modeling techniques that incorporate geometric and mechanics constraints can generate a 3D parse graph for enabling semantically meaningful 3D reasoning [211]  see Figure 35. The 3D parse graph code is available at: http://balaton.graphics.cs.cmu.edu/abhinavg/hn26/blocksworld.tar.gz.

  The DARPA Visual Media Reasoning program is also developing 3D scene reconstruction capabilities from a single image by simultaneously optimizing camera calibration,

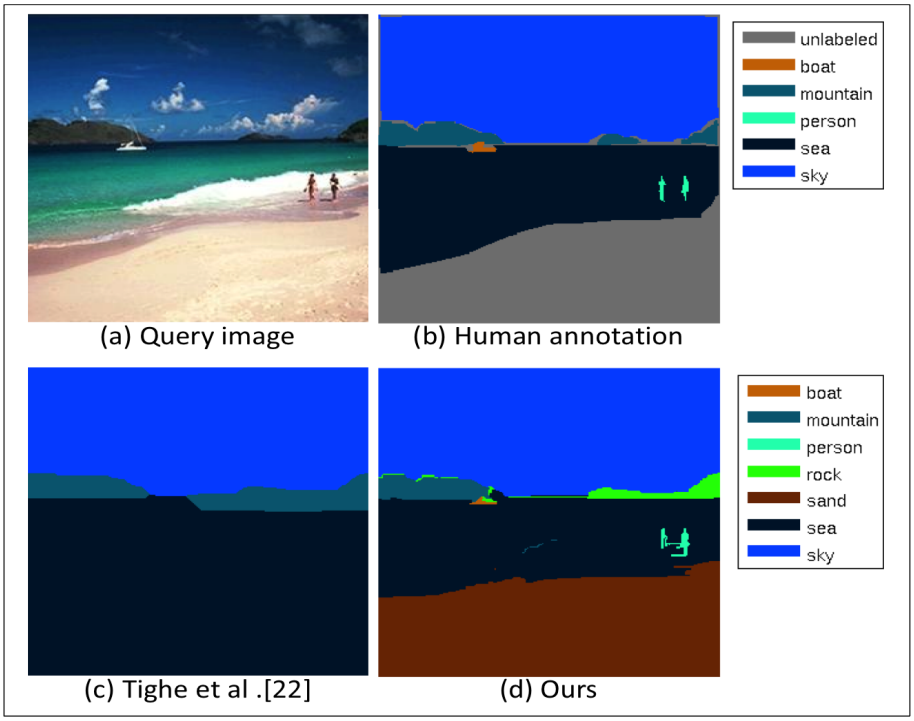



Figure 33: Identifying salient small regions in scenes using an iterative context-driven approach [209].

Figure 34: Make3D estimates scene depth (right) from a single image (left) [210].

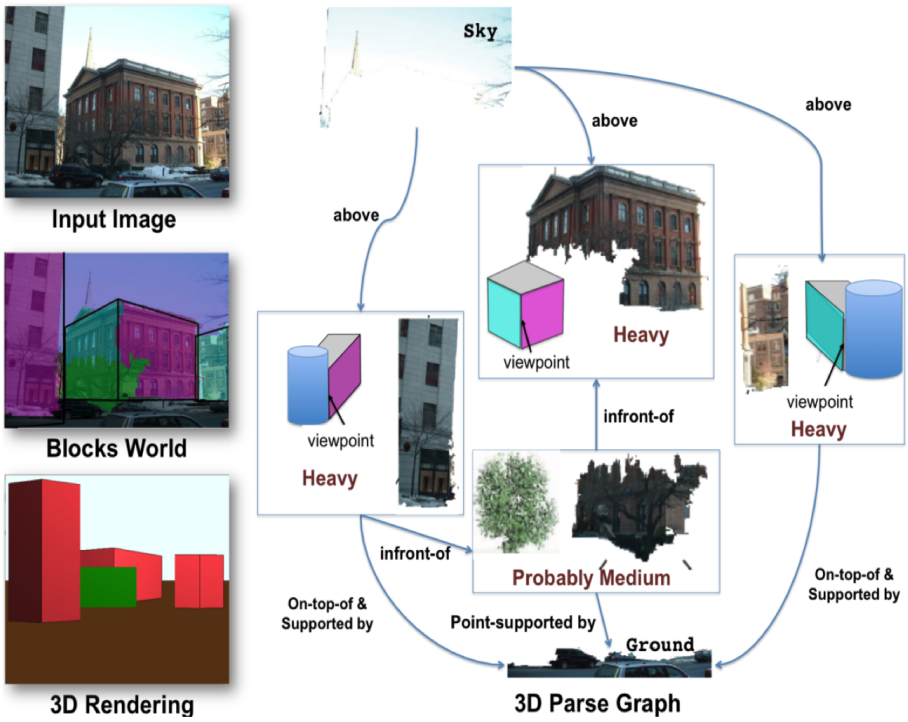



Figure 35: A 3D parse graph captures the 3D relationships between scene entities [211].

scene geometry and metric size variables based on the output of multiple object detectors and image segmentors. Sample output of the estimated 3D locations of vehicles detected in an image is shown in Figure 36.

3D modeling from multiple images or video sequences is an area that has seen extensive development. Many structure-from-¡x¿ techniques exist for a wide range of applications. While earlier techniques relied on accurate camera calibration knowledge, more recent techniques solve for relative image registration and are more robust to uncertain camera metadata. Such techniques include the large scale 3D capabilities of Snavely and Photosynth as well as Probabilistic Volumetric Modeling [212].

- Partially Occluded Object Detection: While detection and classification capabilities of unoccluded objects has improved, even with diverse object types and variable poses, finding imaged objects that are partially occluded remains a challenge. Occlusion may be due to foreground movers, static scene entities, or field-of-view limitations. While some occlusion robustness techniques attempt to predict common modes of occlusion a priori and train for those situations, a more robust solution is needed for addressing the occlusion problem, particularly for cluttered ground-based imagery with moving foreground objects continually obscuring portions of the background objects. Desirable near-term object detection solutions should ideally exhibit graceful performance degradation with increasing occlusion.

  One promising solution for developing improved capabilities is layered reasoning for explicitly ordering relative object depth in the scene. Such a capability, as shown in

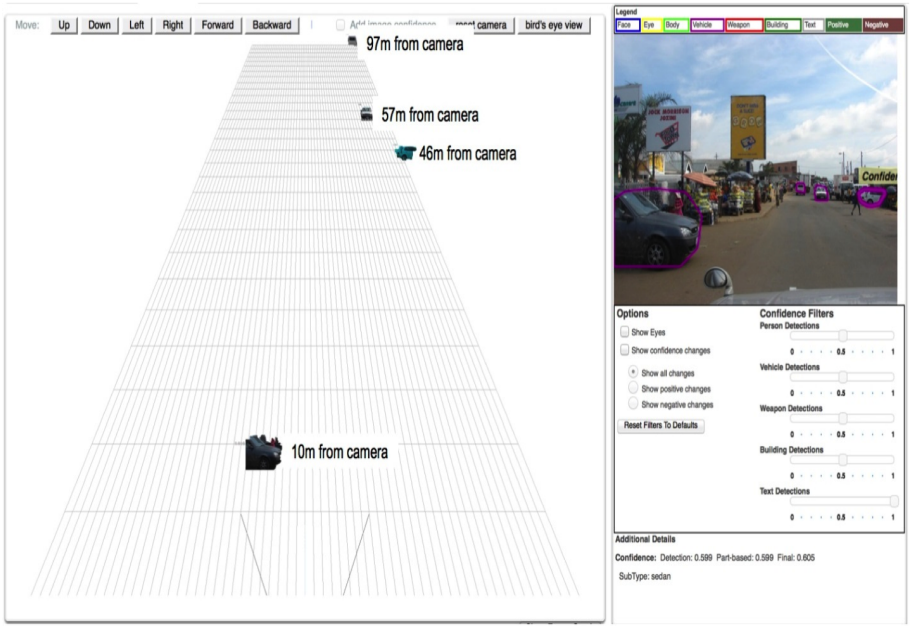



Figure 36: 3D scene localization of vehicle objects detected in an image from the DARPA Visual Media Reasoning system.

Figure 37, can improve background object detection performance as well as improve object segmentation in the scene by separating objects that are hypothesized to be at different depths from each other. An analogous technique is being developed under the DARPA Visual Media Reasoning (VMR) program to apply the scene depth estimates associated with each detected object in a layered detection framework to mask foreground objects from the detection processing of background objects. In VMR, the 3D depth estimates are also used to verify object visibility i.e., ensuring that the depth of small objects place them in front of large (opaque) objects.

- Multi-scale Efficient Semantic Summarization: A growing emphasis in the computer vision community is on the design and implementation of efficient algorithms. Particularly since many computer vision algorithms are computational complex, image analysis capabilities are being developed to leverage algorithmic optimizations and embedded hardware platforms to achieve practical systems. Such efforts are important since camera resolutions and volumes of collected imagery are growing quickly, so the efficiency of solutions must be considered in tandem with the accuracy and robustness of the solutions. Gigapixel images, which are now being generated through mosaicing techniques, will be more common in the not too distant future [214] see Figure 38. A 1.8 gigapixel video camera the DARPA ARGUS-IS sensor has already been developed. A focused investment on efficient multi-scale semantic summarization will ensure practical solutions evolve that fully leverage the rich information available in emerging high-resolution cameras.

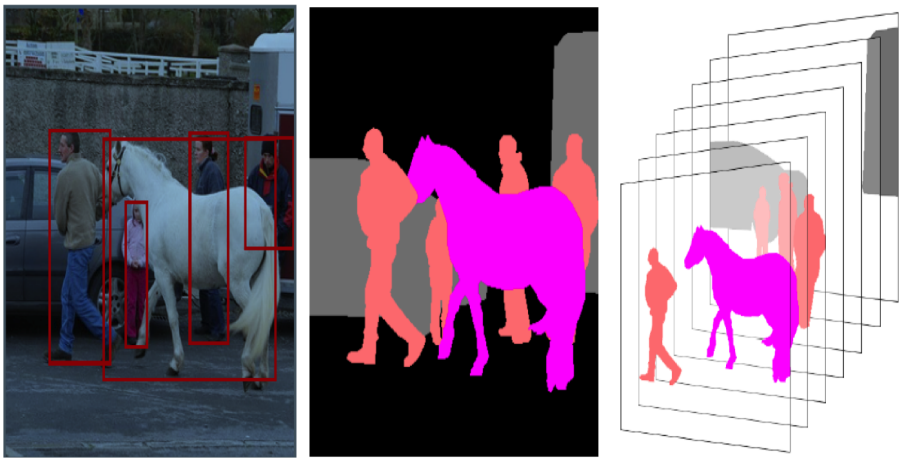


Figure 37: Layered reasoning is a promising solution approach for detecting partially occluded background objects [213].

Figure 38: Sample gigapixel image demonstrating the need for efficient multi-scale image processing [214].

## 8.3 Problems that Need Long-term Investments

In this section, we summarize desirable scene summarization capabilities that will require focused research efforts to develop and will likely take more than three years to mature. Suggested long-term developments are in the areas of:

- Characterization of Dense Interacting Objects: As shown in Figure 38, imagery will increasingly provide higher resolution data over broader fields-of-view to stress the generally bottom-up processing of semantic summarization processing. That is, it will become increasingly necessary to integrate top-down scene understanding to develop interpretations of aggregations of objects along with the conventional processing framework that builds interpretations from pixels to regions to parts to objects to attributes to relationships. New challenges include the direct detection of "groups" and "crowds" and characterizing their approximate counts, distribution of appearance attributes and distribution of poses.

- Functional Labeling of Scene Regions and Objects: Conventional semantic summarization techniques apply a range of feature extraction algorithms to identify scene regions, objects and scene geometry. But semantics are not only based on appearance properties, but also on function such as, what a person is doing with a tool and how boxes on a street are being used (trash can, mailbox, etc.). Datasets are now being developed to enable exploration of functional reasoning, particularly for video sequences. For example, the Penn Functional Scene Element Dataset is a newly curated dataset for functional scene elements containing over 8 hours of annotated HD video for 11 functional classes: bench, bike rack, bus stop, crosswalk, door, news box, parking kiosk, road, sidewalk, subway, trashcan. And initial capabilities are being developed to extract function from video sequences, such as [215]. Research efforts to integrate functional, appearance and geometric reasoning will yield significantly more meaningful semantic summarization capabilities.

- 4D Scene Modeling: A potentially powerful extension to conventional 3D modeling is the integration of spatial and temporal reasoning to develop 4D scene models. Such 4D models will reconstruct 3D models of both stationary and moving scene entities, place the movers in scene in their 3D coordinates as a function of time, track the movers through the dynamic scene, and enable immersive scene inspection from novel viewpoints at arbitrary times. Research efforts to develop solutions for these problems needs to address numerous challenges: spatio-temporal image filtering, spatial and temporal camera calibration, 2D and 3D data registration, spatial and temporal uncertainty estimation and propagation, integrated 2D/3D object detection/classification/matching, spatio-temporal context development and application, novel 4D representations, and efficient 4D modeling computation.

- Anomaly Detection and Intent Recognition: As semantic summarization capabilities mature, they will not only enable human high-level reasoning, but will also set the stage for automated high-level reasoning algorithms. Sample high level reasoning algorithms include anomaly detection and intent recognition. Anomaly detection develops

an understanding for common scene/object attributes so that deviations can be automatically detected. At issue is the need to detect the anomalous behavior in the first place and to avoid high false alarm rates as a side effect. Challenges include the development of normalcy models with possibly limited data and parameterizing the anomaly detection processing with sufficient contextual information to truly separate anomalies from normal behavior. Intent recognition extends spatial attributes of people to include emotions and expressions. Humans read faces very well, but researchers are only starting to scratch the surface of algorithmic approaches to read emotions and expressions.

- Confidence Estimation of Semantic Summarization Results: An issue with semantic summarization, as well as in many image exploitation algorithms, is the need to estimate confidence of the produced results. While the results may never be perfect, can we at least know when we are less sure of the correct answer? The need is to develop performance models that relate operating conditions to performance accuracy. So when a semantic summarization system develops multiple conflicting scene interpretation it will be able to reliably determine which one is likeliest to be correct, or possibly even return multiple hypotheses to provide a user with the most useful information.

# 9 Large Scale Visual Recognition[7]

## 9.1 Introduction

Scaling up is the ultimate goal for all vision tasks, and remains the biggest challenging for system building practice in the history of computer vision and pattern recognition. It is also a major driving force for new representational, computational and learning paradigms. The term large scale has meanings in at least three aspects, which we will summarize in the following.

- Classification and detection with large number ($10^2 \sim 10^3$) of object categories: In computer vision, we often talk about object categories at three levels, as Figure 39 illustrates.

- Entry-level object categories: This is often posed as a classification and detection problem. The method must accommodate the large variations of object instances inside each category.

- Fine-grained and attribute recognition: For example, the type of birds and dogs can be very large. The attributes can be a part of the object or some global style and function of the object. The most challenging and useful fine-grained recognition are for humans and vehicles. The human attributes include the global descriptions, such as gender, age, ethnic group, etc. and the local descriptions, such as hair style, jacket, jeans, etc. The vehicle attributes includes the parts, models and makes. Such descriptions are useful for narrowing the searches in the dataset of security surveillance or large set of images generated by the crowd.

[7]Trevor Darrel, Alex Hauptmann, Gang Hua, Feifei Li, and Song-Chun Zhu

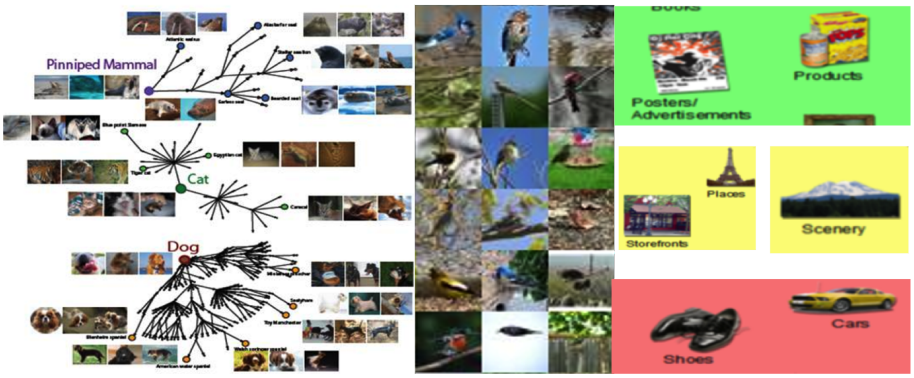



Figure 39: Large scale visual recognition at three levels: (left) general entry level object category recognition [216]; (middle) fine-grained and attributes recognition [217]; and (right) Instance level recognition for products and landmarks. (Images courtesy of Gang Hua).

- Instance level object detection: For example, finding the same products, landmarks in images. This is useful in e-commerce and organizing photo-album.

- Joint image/scene parsing and reasoning: The goal of visual recognition is not merely object classification and detection, but aims for a comprehensive and holistic understanding of scenes in images. These tasks are summarized in Figure 40. The vertical axis represents the objects, scenes and events as increasing complexity; and the horizontal axis goes from syntactic image parsing to semantic image understanding. Image parsing decomposes an image into constituent objects and parts in a parse graph structure and constructs 3D scene layout. Image understanding further augments the parse graph with semantic relations and associates the nodes with attributes, functions, and goals/intents etc. A joint inference of all these sub-tasks will lead to a comprehensive understanding of the scene, which provides the spatial and temporal contexts for recognizing the objects in images.

  During the workshop, participants generally agreed that most vision tasks are not solved, except for a few tasks in isolated situations (e.g. license plate reading) which are deployed by companies with profitable products.

  Even the simplest vision tasks, such as edge detection, cannot be solved or even cannot be well-defined without understanding the whole image. Therefore, it is characteristic of visual recognition tasks that either all of them are solved together or none is solvable. In other words, the lack of such joint inference capability in current vision systems is the main obstacle for scaling up. This has become evident in the current performances of feature-based object classification/detection methods on popular benchmarks which we will discuss in the next section.

- Systems for recognition in large space-time units: The third aspect of large scale refers to the space-time volume in the recognition tasks. Figure 41 illustrates two examples: one in a nursing station and the other a building and parking complex. In these scenes, humans (and vehicle) are tracked across a network of cameras covering a large area and

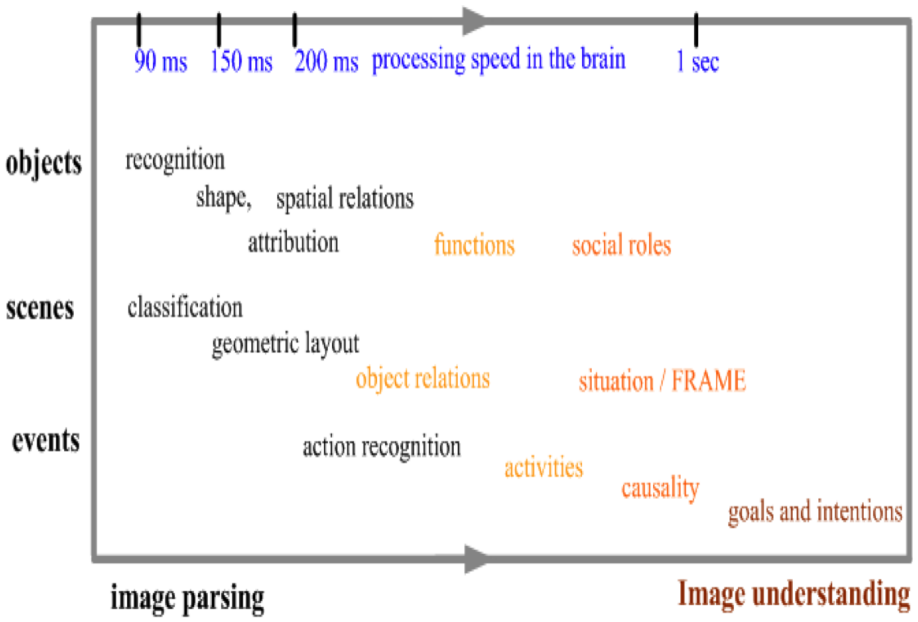



Figure 40: The scope of recognition: from syntactic image parsing to semantic image understanding. This demands a joint solution.

over a long time duration. A meaningful event usually involves multiple people, vehicle, objects, and their interactions. Through these activities, the recognition system can also reason about the scene functions, and predict the intents of agents in the global scene contexts.

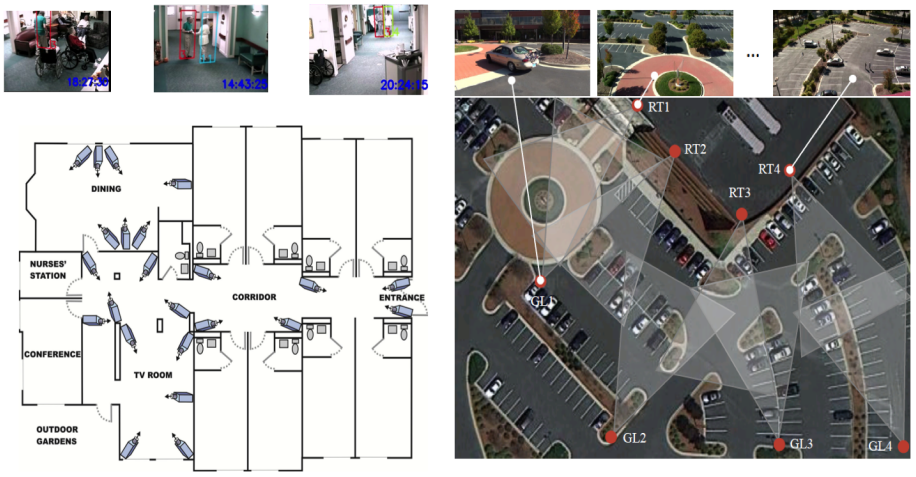



Figure 41: Visual recognition of events, actions, humans and objects in a large area over long time using a network of cameras, (Left) a nursing station with 23 cameras (Courtesy of Alex Hauptmann). (Right) a building and parking complex in a DARPA MSEE project (Courtesy of the Signal Innovations Group).

Besides the recognition tasks, there are other technical issues, such as time synchronization, view registration and 3D scene reconstruction etc. All of these are the necessary functions for building surveillance systems and for enabling autonomous robots to navigate in such complex environments.

Again, scaling up is a major technical challenge for commercialized applications of such systems. Here the concept of large scale also includes the aspects I and II above.

## 9.2 State of the Art

Despite the vast literature on visual recognition, the underlying methods used in research can be roughly divided into two streams.

- Feature-based SVM classification. Methods in this stream extract features, arrange them in a very long vector, and feed this vector to SVM training for a one-vs-other classification. The SVM allows non-parametric weighting of the features. The features can be diverse depending on the underlying objects and resolutions. Figure 42 shows the architecture of the deep convolutional neural networks (CNN). It extract features through 5 layers of sum- max operations, which leads to a 4096 vector with 2 layers of full connections. CNN makes the features themselves non-parametric and trains the massive number of connection weights through back-propagation minimizing the classification errors. The CNN algorithm currently shows good promise on the ImageNet challenge, and becomes a hot research topic with more results reported in vision conferences.

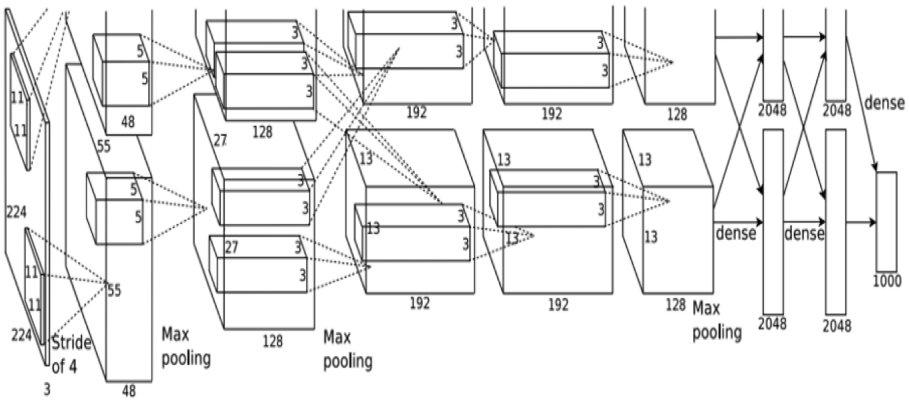



Figure 42: The architecture of the 5+2 layer Deep Convolutional Neural Networks. From [54].

When multiple labels are needed for an image, then one SVM is trained for each label. Figure 43 illustrates the pipeline used for classifying human attributes, such as male/female, wearing hat/not, wearing short/not, long/short hair, etc. In scene attribute classification, for example, trains 101 SVMs for 101 attributes. In summary, the methods under feature-based SVM classification have two characteristics: 1) using

features for SVM training in a feedforward way; and 2) training labels separately and thus the relations between these labels are not modeled explicitly.

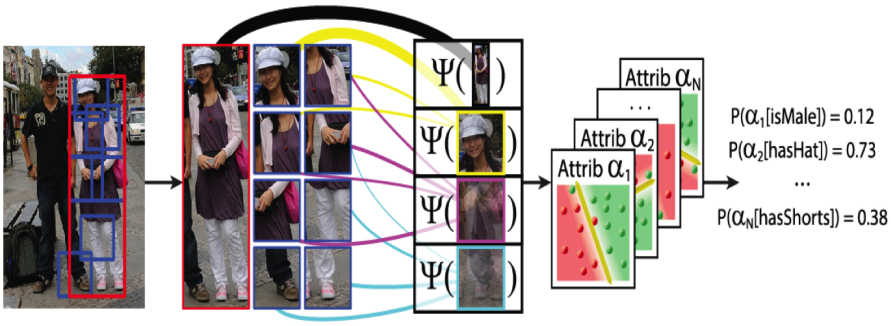



Figure 43: An example of human attribute classification and detection by Deformable Part Descriptors (DPDs). (Courtesy of T. Darrell) from [218].

- Compositional approaches and top-down/bottom-up parsing. In contrast to the SVM-based methods, compositional approaches and top/down parsing strategies aim to parse the whole image and output a hierarchical parse graph. The nodes in the graph represents levels of concepts, the vertical links represent decomposition (part-whole) relations, and horizontal links represent spatial, temporal contextual relations. Instead of computing each node separately, they are inferred jointly, very often, through iterative bottom-up/top-down process. Figure 44 illustrates a scene parsing example by [219]. The parse graph includes i) scene as the root node, ii) compositional objects, such as cube, nested frame, cabinet, floor tile patterns, as intermediate nodes, and iii) rectangles as terminal nodes grounded on detected edges.

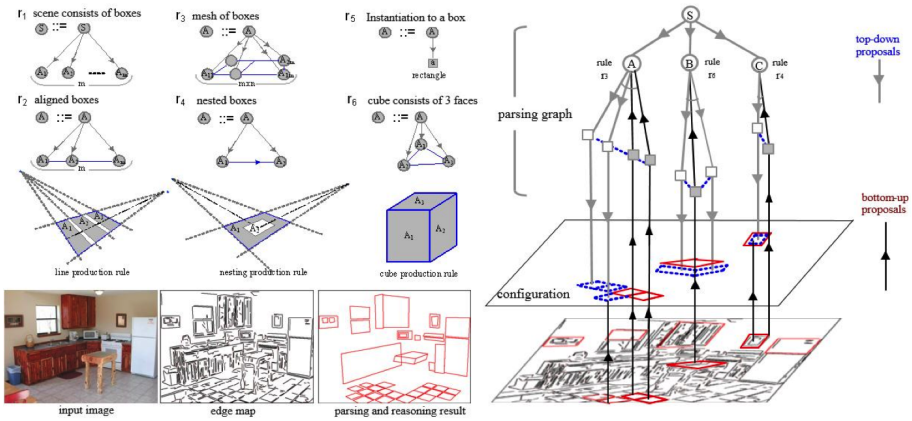



Figure 44: Image and scene parsing with bottom-up and top-down inferences. (top-left) the typical image grammar for the composition of rectangles in line, mesh, nesting and cube. The edge maps is extremely noisy and therefore the algorithm must rely on the top-down and bottom-up iteration (right) to propagate the information between the rectangles [219].

The parsing methods follow the stochastic image grammar and have been recently extended to i) object recognition by computing the scene object-parts-primitive hierarchy; and ii) event understanding by computing the event-action-object hierarchy.

The two streams of methods are not necessarily exclusive to each other. In fact, a strong detector or classifier can be helpful for detecting any node in the parse graph. Therefore an integration of the two streams will likely lead to new state-of-the-art in future.

## 9.3 Solved Problems

For large scale visual recognition, the following functions have been widely deployed in commercial products and can be viewed as solved problems.

- License plate detection and reading: Automated license plate reading was deployed early at restricted environments, such as entrance of a parking structure or entrance of tolled highway. Then as high-resolution video cameras become available, it is used on street and bridges to monitor traffic and control vehicle flows. For example, in some cities, vehicles with certain plate numbers are limited to using bridges and roads at certain days or specific time windows. Figure 45 is an example of a system counting vehicles and detecting their plates at higher resolution, as they pass the camera.

Figure 45: Traffic flow monitoring and license plate reading in real time [220].

- Face detection: This is a function widely used in all digital cameras, smart phones, and in other software for organizing photo albums and social network websites (see Figure 46).

- Pedestrian detection: Pedestrian detection on street is currently quite reliable and is used for security surveillance systems and for driver assistance.

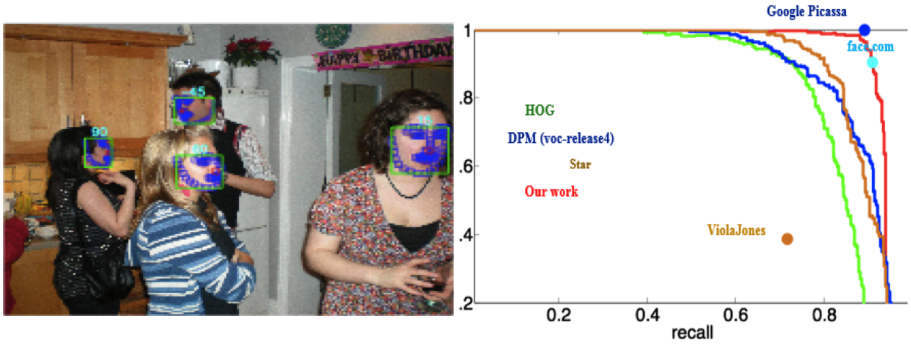



Figure 46: Detecting human faces of varying poses and scales from large scale images in real time [221].

- Near duplicate instances detection: The instance level object detection is useful in near term. This may have been discussed in other workshop topic in details.

## 9.4 Near-term Solvable Problems

In middle term ($3 \sim 5$ years), we expect that the following functions will become increasing reliable and useful in practice.

- Detecting vehicles in general conditions and under severe occlusion conditions: With high resolution cameras or PTZ cameras, this function can be used for reading license in far distance (500m). This is potential useful for managing traffic and providing parking information in big cities.

  A further step that may be reachable and useful is to classify the models and makes of vehicles.

- Human attributes in video: We often describe other people by attributes. There are global attributes, such as gender, age, ethnic group, and professions etc, and local attributes associated with body parts and accessory objects, such as wearing hat, sunglass, long hair, jean, shorts etc. A person also has social dimensions showing expression and characters which are particularly useful for analyzing multi-media data for social and political sciences.

- Human pose estimation: Detecting the body parts of human figure: head, torso, upper/lower arms and upper/lower legs. This is useful for understanding human action and activities.

  Figures 47, 48 and 49 present examples of vehicle detection, human pose estimation and attribute extraction for human using standard data sets, respectively.

## 9.5 Problems that Need Long-term Investments

We expect to see tremendous progresses in the long run ($5 \sim 10$ years) for large scale visual recognition tasks, which we will put in three aspects as we outlined in the introduction.

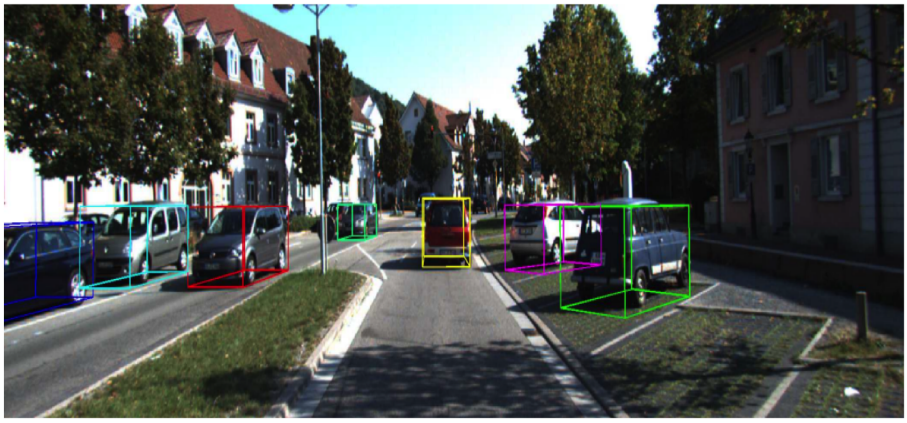


Figure 47: Vehicle detection in the wild. Results on the KITTI car detection benchmark which is divided into three subsets hard, moderate and easy. The Average precision for detection in the KITTI car dataset is about 80% for easy cases, and 50% for hard case. (Courtesy of Bo Li).


Figure 48: Results of human pose estimation on the UC Irvine PARSE benchmark (first row) and Leeds benchmark (second row). The average precision (AP) for head and torso are near 99%, while the arms and legs are about 70%. Some typical failure examples are shown in the 3rd row. From [222].

Figure 49: Human attribute recognition and describing people by their attributes. Examples from the Poselet dataset: male (top) and female (bottom). The current performance of male and female classification reaches about 90% in this dataset [221].

- Large scale object classification and detection:
  - 1000 category object classification and localization for image retrieval purpose.
  - 100 category human action and activity recognition in conjunction with human pose estimation, concurrent action recognition (multiple actions in parallel).
  - At instance level, computer vision will have to solve the long challenges on face recognition [223], human identification and re-identification from video under reasonable conditions.
- Scene parsing and reasoning with commonsense knowledge:
  - From image parsing to image understanding. The integrated tasks were outlined in Figure 40, and Figure 50 shows a concrete example. On the left is traditional image parsing which outputs a syntactic parse tree organizing the objects, surfaces, and primitives in a hierarchy. On the right is image understanding which augments the parse tree by reasoning the functions and possible human actions with the objects. These reasoned activities, in fact, define the functions that the space serves — the essence of scene category, and assign meanings and semantic labels to the objects in the scene. Further augmented properties include physical relation, such as supporting relation between objects.
  - Joint object-scene-action/event inference in video. As we argued in the introduction, visual recognition tasks, such as object detection, human pose estimation, vehicle detection, attributes, action recognition, and scene parsing, should not be studied in separation, and must be solved in a joint inference framework. Figure 51 show a screenshot of the joint spatial, temporal and causal parsing framework with a link to online demo video on Youtube.

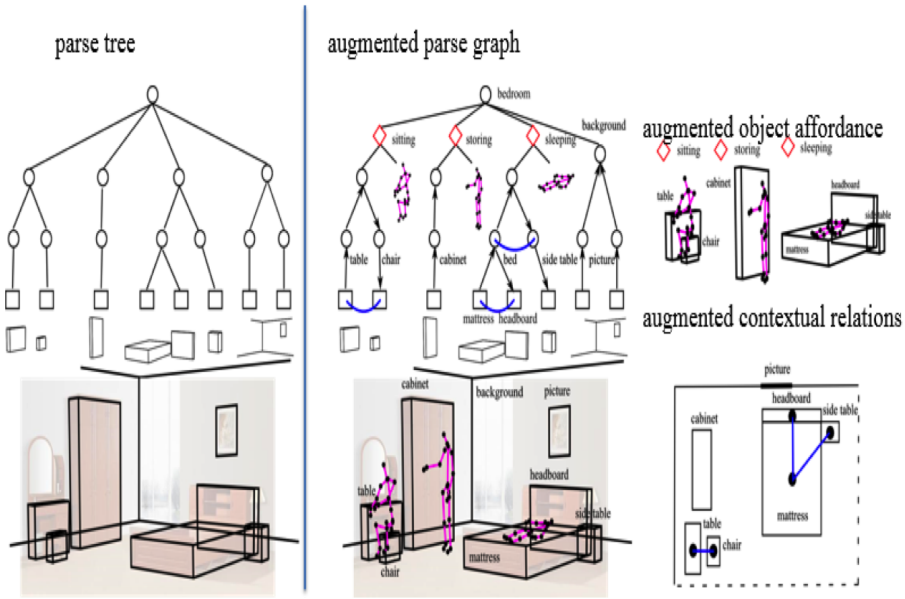



Figure 50: (Left) Image parsing which organizes objects, rectangular surfaces in a parse tree based on their geometric properties. (Right) Image understanding augments the parse graph with functions and and activities as well as various semantic relations. Modified from [224].

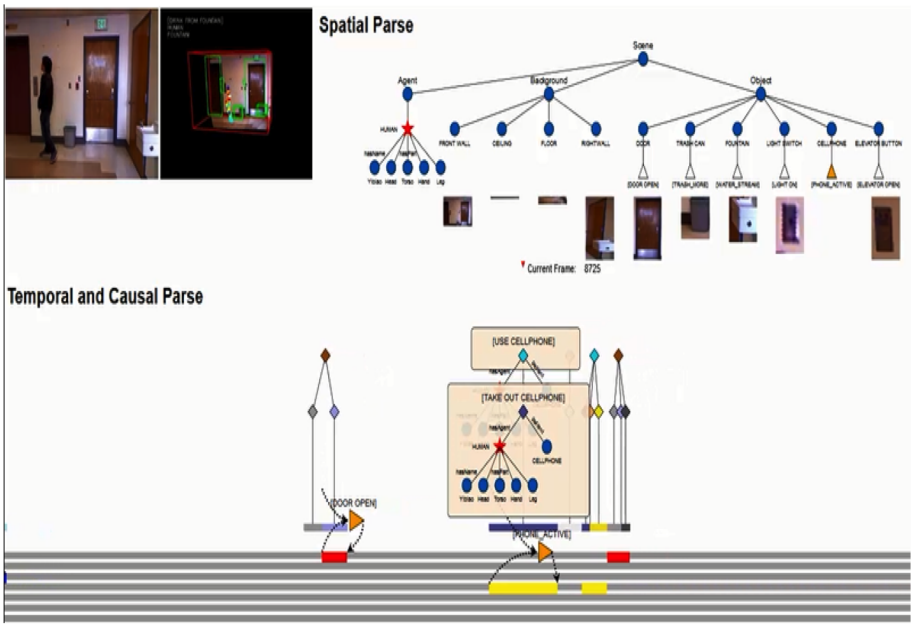



Figure 51: Screen shot of a joint spatial, temporal and causal parsing system for video understanding. The demo can be viewed on Youtube: https://www.youtube.com/watch?v=FmFK52WwSQg.

Recognition is only a marginal task of parsing. As a single image or a video clip contains all these visual concepts, all the recognition problems either can be solved together or none of them can be solved. This observation also leads the vision researchers to study the representations and models of commonsense knowledge. This calls for collaborations with language, cognition, and AI.

# 10 Data Sets and Performance Evaluation for Research in Large-Scale Video and Image Content Analysis[8]

## 10.1 Introduction

In previous sections, we identified tasks in image and video analysis as belonging to one of three distinct groupings. Grouping number one, "solved" problems. Grouping number 2, problems where significant advancements may be anticipated in the coming 1 to 3 years. Grouping number 3, problems where significant advancements in all likelihood lie more than 3 years down the road. A key question for the discussion in this chapter is how to relate these groups to data sets and performance evaluation protocols

One can identify a "solved" problem based upon the fact that good data sets for evaluating performance on that task have already been in use by the research community for some number of years. Similarly, if we want to see documented progress on a problem in the 1-to-3-year time frame, then the relevant datasets should have already been collected, and either already be available or be about to be made available to the research community. Perhaps most importantly, any task for which there is not already a good data set introduced to the research community almost has to be considered as lying beyond the 3-year horizon.

Given the diversity of computer vision topics discussed in this report, general statements about data sets and evaluation protocols covering these distinct areas are problematic: different tasks impose different needs and constraints. To begin making sense of the range and scope of tasks, it is helpful to think about 3 kinds of tasks: tasks involving places, tasks involving objects, and tasks involving human activities. Obviously, this taxonomy is simplistic, e.g. tasks may involve questions about people doing things with objects. However, it is far better to start with this taxonomy than with no taxonomy at all, and further this taxonomy reflects the reality that most current data sets may be easily associated with one of these three categories.

The design of performance evaluation protocols and data sets largely depend on the task being addressed. For example, research into how to build systems that solve geo-location problems has exploded in the past decade. One particular variant on this task is to answer the question: "What were the GPS coordinates of the location of the camera when this picture was taken?" In terms of data collection and metadata, this particular task may be an anomaly as this metadata comes for free using most modern cameras. However, as soon as the task definition shifts to questions such as: "What are the positions and orientations in 3-D of the 4 cameras that viewed the following scene on the Washington Mall?" The second question, while also very much focused on questions involving places and space, implies an

[8]Ross Beveridge, Kevin Bowyer, John Garofolo, Jonathon Phillips, and Ranghachar Kasturi

entirely different and much more demanding set of constraints with respect to evaluation and data sets. Put simply, the necessary grantors and metadata does not arise easily for such tasks.

Acquiring millions of images off the web of "things" is spurring a different explosion in research and development. Unlike geo-location information that comes for free in many contexts, labeling images with their contents, i.e. airplane or backpack, still requires some human effort. That said, it appears to be a task that can be addressed through various forms of crowdsourcing. Large data sets have already been established and larger data sets are coming online. It should be noted that the degree of ground truth metadata that can be constructed after the fact remains a somewhat open question. Another issue with these data sets is ownership and how data is distributed: at the moment some of these data sets represent collections of URLs rather than data.

The last category, tasks involving people, bring along a set of particular complications not necessarily shared by the previous two tasks. The research community as a whole is sensitive to issues of privacy and more generally the appropriate safeguards governing the use of data associated with people. It is also the case that when doing research into tasks involving people doing stuff metadata regarding what they are doing, even if it is simply looking directly at the camera, is essential. Data sets that lack accurate information about identity and the conditions under which the data is collected are relatively diminished value.

Data alone is of little value. Value to the community rests upon three things: 1) publicly available and well organized data, 2) a clear evaluation protocol with metadata and explicit experiments, and 3) software support including examples of how an experiment is run. One term used to capture all three taken together is "Challenge Problem". Challenge Problems are most useful for advancing technology when they capture shared aspects of multiple operational tasks. For example, the FERET collection was not designed to match any single operational scenario, yet the resulting data, evaluation protocol and benchmark algorithms spurred research and development for a range of real-world applications. Going forward, the best data sets should likewise lie one step removed from any particular operational scenario while at the same time capturing essential elements of multiple operational tasks.

Finally, given the plethora of evaluations and data sets that have permeated the computer vision community over the last two decades, it is clear that technology advances when evaluation protocols emphasize understanding of the task and provide diagnostic information. While it is natural that attention is drawn to top-performing algorithms, it is critical to establish a context for interpreting results and a means for the community to understand and improve performance based upon diagnostic information explicit in those results. It is also important that community reaches a method for retiring a dataset and a challenge.

## 10.2 Role of Baseline Algorithms, Datasets and Performance Evaluation in Accelerating Research

The connection between research progress and the availability of data sets, baseline algorithms and evaluation protocols cannot be overemphasized. A fully supported data set that includes appropriate human generated annotation, a clearly identified evaluation protocol, and openly available baseline algorithms will go a long way in contributing to research

progress.

A profusion of datasets are already available with more arriving every couple of months. Typically, each occupies a particular niche with its own application focus and/or set of particular attributes with respect to content and construction. In the context of video, some data sets emphasize basic human motion, others emphasize explicit human activities and even others emphasize large numbers of people viewed in crowds. In some cases actions are staged, in other cases actions are captured in public settings. Whether objects play a role varies widely, as also the type, size and nature of object. Along an entirely different dimension, how video is captured varies greatly. There are data sets collected using single cameras, multiple cameras with distinct views and multiple cameras with overlapping views. Some data sets are collected outdoors, some are collected indoors, some have a mix of the two. For the research community today, it is not too strong a statement to say that progress is being hampered by the ready availability of raw data and the degree to which this is pulling the community away from a focus upon common data sets.

As the cost of acquiring imagery and video continues to drop, niche data sets will in all likelihood continue to proliferate. This fragmentation comes at a cost. For so many of the tasks associated with image and video analysis, the only means of assessing quantitative progress is through carefully specified experimentation on common data sets. It is imperative now, just as it has been in the past, to continue to try to promote research community investment in common well-designed open access data sets. This need to encourage research back towards common data sets and protocols is timely in the face of the rapid advancement of cloud computing. Cloud-based open access data sets and associated evaluation support framework should be seriously considered and work begun on their construction.

Annotating a data set is the process of recording as "ground truth" essentially any aspect of the associated scenario that might be considered important. Often, the results of the annotation processes are referred to as metadata. At the most basic level, annotation establishes the names of the objects and object recognition data set, or the identities of the people in a face recognition data set, or even perhaps the GPS locations of photographs in a geospatial data set. In rare cases, metadata is required at essentially no additional human cost as part of the process of capturing images and video. This is true for some low-level camera supplied data such as f-stop setting. It is also increasingly true that GPS coordinates are automatically inserted into digital imagery, thus geo-tagging of data is to some fair degree the exception to the rule when it comes to the amount of human labor required to create metadata.

Unfortunately, for a great many other tasks, annotation still involves human labor. That labor can be expended up front as part of the data collection process. Alternatively, in some cases it can be done after the fact by review of the data. Annotation can even in some cases be done through crowdsourcing of data after the fact. As a general rule, annotation anticipated as part of an experimental design and data collection procedure will involve less human effort than after-the-fact annotation appended to data by post-hoc human inspection. While annotation is an expensive procedure in many cases, it plays a critical role in the systematic evaluation of end-to-end prototype systems and is an essential component of data sets expected to truly promote advancement. Open access to readily understandable and recognized baseline algorithms paired with a data set greatly increases the value of a data set. One way in which baselines are important is they reduce the amount of time

distinct research groups spend "reinventing the wheel". There are at least two ways this is true. In broad terms, a well-recognized baseline helps delineate what might be considered the solved parts of the problem and spur researchers to move away from minor variations on prior art. The second is more detailed and pragmatic. Often, by studying the parts of common baselines, it is possible to identify and reuse some components while breaking new ground with others. Finally, quantitative performance results from baseline algorithms let developers soundly judge under what conditions state-of-the-art solutions are sufficient to solve their specific problems, and also, to correctly judge when additional basic research is needed.

One must draw a contrast between evaluation protocols focused on ranking performance versus those designed to provide diagnostic information about performance. Simply measuring the difference in performance between two alternative algorithms by itself provides relatively little guidance to the research community regarding where to focus energy and creativity. In contrast, when the research community at large collaborates and agrees on metrics and particulars of evaluation that provide direct diagnostic feedback, then it is possible for the development process to rapidly advance in response to that feedback. One of the most important single things that can be done to advance research in image and video analysis is to make sure that an open evaluation infrastructure including data, standardized algorithms, and well-defined evaluation protocols is promoted and adopted by researchers.

## 10.3 What Makes a Good Data Set?

The first answer is the amount of money spent to create the data set; this is an option that is relatively easily dismissed. A slightly more sophisticated way to measure the value of the dataset is simply count the number of groups using it in their research. However, the weakness in a simple counting argument isn't hard to imagine: popularity does not equate directly with significance and impact. A variation of this counting argument might be to count number of publications using the dataset, but this suffers from the same sort of problem. Last, and by design hopefully the most compelling, the value of the data set lies in the value of the science, engineering & policy advances made possible by using the dataset. But, as the sequence of questions suggests, as the criterion for assessing value becomes more compelling, so does the difficulty in quantifying and objectifying that standard.

In the context of collecting and distributing data, the issue of appropriate access policies for data sets has to be addressed. There is a broader spectrum of policies already in play with existing data sets than might at first be apparent. At one end of the spectrum lies unrestricted download from the web. At the other end, lies access to data only in secured facilities. In between these extremes lie several different degrees of control over access. For example, data sets may be available to researchers worldwide contingent upon a signed legal agreement. Data sets may be available to a restricted set of users. Data sets may come with restrictions on how the data is managed and stored. Finally, some data sets are available only after endorsement of the user by a representative of a government agency. The one high-level conclusion that is perhaps obvious is that every obstacle placed between a research group and a data set reduces the odds that data set will ever actually play a significant role in advancing research. Datasets that can be used only on signed agreement that they will never be stored on a computer connected to the internet are unlikely to see much use.

Given the recent trend in collecting and experimenting with "in the wild" data sets, it is useful to evaluate the value of such data sets. At the outset, the meaning of the phrase "in the wild is vague. Conventional wisdom suggests that research requires data sets to be representative of some specific intended application or set of applications. Appending what has now become the adjective "in-the-wild" to a data set does not always reveal the intended applications or the underlying assumptions, constraints and nature of a data set.

Another issue is concerned with the value of "operational" data. It's not hard to understand why, at face value, the notion of real operational data, pulled perhaps from log files of fielded systems, would be intrinsically more relevant and useful than data collected in a carefully designed, managed and scripted data collection protocol. However, making such a reflexive leap is not always desirable as it overlooks very real constraints implied by operational data. First, an operational system must achieve a threshold level of success to continue to exist, and this alone implies a great deal about the nature of the data it will acquire. Second, operational systems must satisfy cost and efficiency constraints that often are at odds with data screening, ground trothing and retention. Third, privacy concerns also may impose limits on data retention. Finally, critical metadata likely will not be collected unless there is some specific, operationally cost-effective reason to do so, and more often than not such incentives are not present. In contrast, research data collections are designed explicitly for the purpose of answering research questions, and the best data collections typically are undertaken only after those questions have been explicitly identified. The conclusion is that while in some circumstances operational data may be of tremendous value, all of the caveats on associated issues just outlined must be explicitly addressed before recommending evaluation on operational data. Going forward, it is important to entertain ways of collecting data that combine the better elements of the two approaches. That is, a specially-instrumented collection in an operational scenario, or a research collection more explicitly designed to mimic important elements of an operational scenario.

## 10.4 Examples of Effective Data Sets

As noted above, many useful data sets have been collected for performance evaluation of various computer vision algorithms. The FERET data set is probably one of the first data sets that was widely used in the nineties for evaluating face recognition algorithms. One can point to three reasons that made this data set very useful. First, this data set was first designed to serve a purpose and then collected. That purpose was to test: "How well can algorithms reliably recognize faces under favorable conditions?" These conditions being, the person is well lit, they are typically looking at the camera, and the background is not distracting.

Second, the FERET data set came with a well-defined evaluation protocol that others could replicate and, which included a small but useful set of sub problems designed to be a graduated difficulty. In very pragmatic terms, one consequence of that design decision was that even as the easiest portions of the data set ceased to be interesting in terms of measuring algorithm performance, or challenging parts of the data set remained challenging well into the 2nd decade of the data sets life, and there are indeed even today labs publishing results on the very hardest remaining portions of the FERET data set.

Two other closely related data sets were collected according to a careful design to serve

a very specific scientific purpose. These are the Yale B and the PIE (Pose, Illumination and Expression) data sets. Both of these data sets varied facial illumination according to carefully varied laboratory conditions. Focusing specifically on the illumination aspects of these two data sets, they were created at roughly the time that a fundamental theory was developing with respect to object illumination, and the tight coupling between the theory and empirical sides of better understanding of illumination meant that these 2 data sets played a pivotal role in the development of the underlying science of facial illumination variation.

A series of data collections carried out at the University of Notre Dame has resulted in face and video data collected according to a plan and distributed with associated metadata. One major example to come out of this effort is the Face Recognition Grand Challenge data set. This data assisted in advancing face recognition from frontal face images collected with digital single lens reflex cameras and 3D scans of faces. In terms of sheer utilization by the research community, one of the most significant data sets of the last decade for face recognition is the Labeled Faces In The Wild data set. This data set was assembled at the University of Massachusetts by collecting images of celebrities off the World Wide Web. The data was compiled into a single repository hosted by the University of Massachusetts. Anyone in the world can go to this website and download the data.

To the great credit of those who put together the LFW data set, it introduced new energy and life into the research community. The immediate apparent difficulty of the LFW data set combined with its obvious real-world origins combined to spur on new research and a variety of important publications. In the past decade, the computer vision community has made tremendous efforts in constructing numerous datasets. For example, the Lotus Hill project hired full time employees to annotate and parse millions of images and videos, and the MIT LabelMe project attracted volunteers to segment and label large number images of scenes. Both projects intended to let researchers select subsets of annotated images and define their own tasks and benchmarks. There are many other more specific datasets at rather large scale for specific object classes, such as face, pedestrian, vehicle, action etc. It is not exaggerating to say that datasets and benchmarks have steered vision research in the past decade.

An interesting observation is that the most popular datasets are not the ones with comprehensive parsing and labeling, but the ones with simply defined tasks — classification and detection. The well-known examples are the Caltech101 classification benchmark, and then the PASCAL VOC object detection benchmark, and most recently the ImageNet benchmark.

The PASCAL VOC detection dataset has 20 classes: aero, bike, boat, bottle, bus, car, motorbike, train, bird, cat, cow, dog, horse, sheep, pedestrian, plant, chair, table, sofa, TV. This most competitive and popular algorithm for this dataset is the deformable part-based model (DPM) [3] coupled with the HoG (Histogram of Gradients) features. The average precision (AP) is below 35%, with functional categories, such as table and chairs having much lower detection rate. The PASCAL VOC benchmark competition ended in 2010. Figure 52 and Figure 53 show some examples from the PASCAL VOC and ImageNet challenge, respectively.

The ImageNet collects a huge set of image categories (1000s) and labeled them in a hierarchy through Crowdsourcing using Amazon Mechanic Turk, but not all categories are used in the benchmark which contains 200 hundred categories. An image often contains multiple objects at diverse scales which provide mutual contexts to each other. The ImageNet

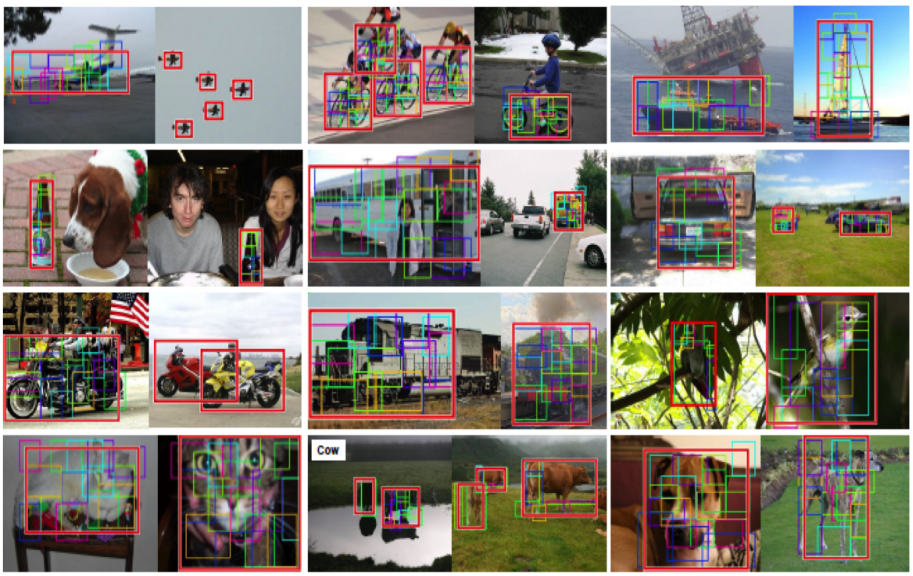



Figure 52: PASCAL VOC 20-object class detection. Each object is detected by a bounding box, and the accuracy is measured by how well the detection box overlaps with the human labeled box. The red boxes are detection windows and the smaller boxes are the parts of the object. (Courtesy of T.F. Wu)

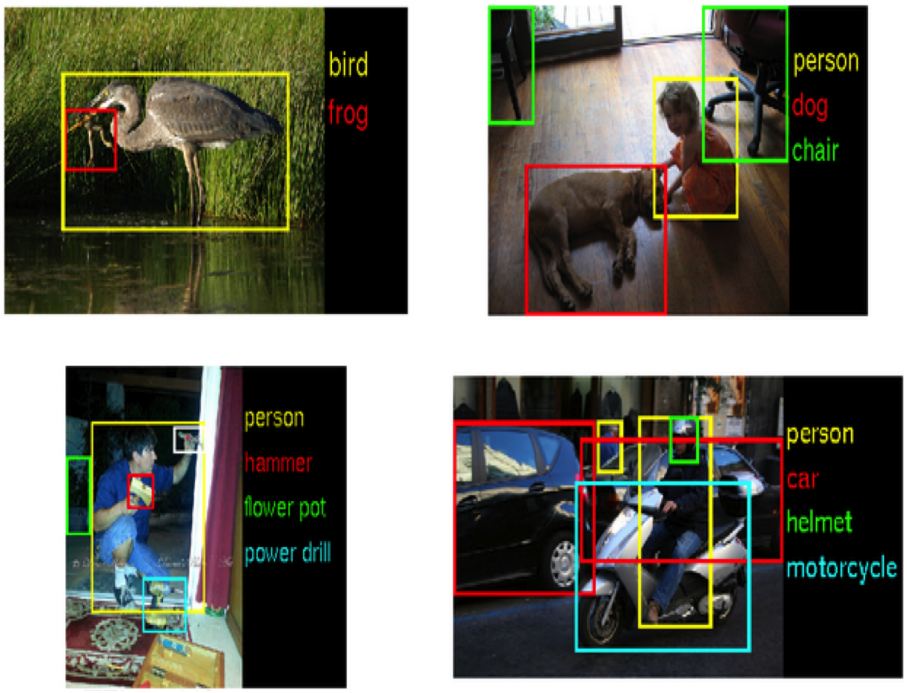



Figure 53: ImageNet Large Scale Visual Recognition Challenge 2013: 200 classes, 395,909 training; 20,121 validation; 40,152 testing. Each image contains multiple objects.

benchmark and competition contributed to the popularity of the DeepLearning method, especially the Multi-layered (7 layers) Convolutional NeuralNet (CNN). The variants of CNN has also demonstrated the much improved performance on the PASCAL VOC 20 class detection to over 50% [4], though it was pertained using the large ImageNet dataset.

## 10.5 Perspective in Directions in Video Analytics Evaluation Challenges/Opportunities

It is clear that the amount of video data being collected is growing exponentially in both consumer and commercial applications. Also, cloud and related scalable architectural models are further accelerating this dramatic up-scaling of data and the associated demand to interpret data. Against this backdrop of exponential growth it appears that analysis technologies are still fragile and true semantic understanding of video is beyond the scope of current technology. Furthermore, the current state-of-the art is primarily focused on individual video stream analysis. Only loose couplings of analytics are employed when multiple cameras are in use. Integrated scalable situational analytics are comparatively weak.

Several significant technology trends and anticipated growth areas in video analytics can be envisioned, starting with spatial analysis of large areas. The key idea here is to provide access to video surveillance in terms of where an event is taking place. To be clear, "where" in this context refers to a physical location, not an individual camera: excessive reliance on humans equating cameras to physical locations is a consequence of a lack of spatial analysis in many currently operational systems. The shift from camera focused analytics to spatial analysis requires independent modeling of a physical location, potentially incorporating 3-D modeling incorporating dynamic calibration of the relationship between camera parameters and physical space since cameras may move and PTZ camera views will change. This shift from a camera to a spatial approach to analytics is harder than may be apparent to those not familiar with the underlying technology. Significant progress should be considered a basic research as well as a development problem.

The next growth area is temporal analysis. If spatial-analysis addresses questions such as: "Where on this map are people exchanging packages?", then temporal-analysis addresses the related question: "Show me all the times that people exchanged packages?" There is a spectrum of tasks in this general area. They range from relatively simple, e.g. event triggering when a person enters an off limits area, to complicated questions involving time and behavior. For example, consider this query: "Show me all times a person stops and peers into a room through a half open door."

Spatial analysis and temporal analysis are not independent tasks since real world queries of a surveillance system will typically involve elements of both. As these two growth areas develop together they will spur basic research on how to create smarter sensor networks as well as ways of constructively fusing observations in support of high-level interpretation. Again, to illustrate the range of problem complexity in this context, a task not far beyond the current state-of-the-art might be exemplified by a query such as: "Show each time a person sets down a package on one of the following tables?" A much more challenging task with respect to fusion over multiple cameras and sensors would be the following: "Show me the pattern of movement of a person over time" where the views of a person across multiple

cameras and space support neither continuous biometric recognition techniques nor motion trajectory tracking techniques.

The last area for growth that is sometimes overlooked concerns the need for entirely new methods of interacting with video and the results of video analytics. As was discussed elsewhere at this workshop, video analysis systems exist to serve the needs of human seeking answers to questions about what has transpired in a physical location over a time window. The graphical user interface issues associated with providing people with just what they need, and not more, are critical. This broad problem of what to present to people and how to present it should be viewed both as a research and a development problem. Significant research needs to be devoted to the effective presentation of video and visualization of video-derived information to minimize both cognitive overload and burnout as humans are not effective at continuous detection of objects and activities in longitudinal video analysis tasks.

How information is presented as a question cannot be divorced from the related question of what current analytics can automatically extract and interpret. The situation is a classic chicken and egg problem. Design of a useful user interface is heavily constrained by what the automated procedures accomplish. Conversely, what video analytics should be accomplishing needs to be driven by what users need. Video analytic technology developers also need to begin to view video as one sensor modality in the context of a suite of potential sensor inputs and create integrative approaches to analytics interfaces.

## 10.6 Video Analytics Challenges for Evaluation

Against the backdrop of an exploding video analytics market, the challenges that face anyone seeking to create data sets and usefully evaluate progress come into stark relief. In particular, the challenge might be summarized as that of creating evaluation "environments" representative of real at-scale analytic challenges. Such evaluation environments need to include either real operational data or realistic data of a kind similar to that arising in an operational context. Further, and this is typically where the greatest expenses arise, that data must come with ground truth annotations sufficient to quantify performance along all of the dimensions identified as being relevant to the task.

A precondition for the design of a data collection and evaluation viewed in this light becomes a concrete specification of how information will be processed and the overall information processing framework associated with some particular domain. Arriving at such a specification involves making explicit the tasks that are to be performed. Further, invariably decisions have to be made about piecewise versus integrated evaluation. An example of piecewise versus integrated evaluation arises in the context of person tracking. Following the model that evaluation needs to quantify performance accuracy of key system components, quantified comparisons between alternative tracking algorithms based upon carefully derived ground truth becomes a high priority. The implication is that having this explicit knowledge will promote the choice of better tracking algorithms and consequently improve those video analytics dependent upon tracking. Furthermore, scalable methods need to be developed to assess the performance of tracking over large spaces and temporal windows. Tradeoffs between painstakingly-created manual ground truth with high precision and technology-leveraged ground truth at lower levels of precision become significantly important at these scales.

In contrast, the integrated evaluation model downplays the need to quantifying tracking as an explicit performance task in favor a higher-level performance characterization. For example, recording the number times a system correctly reports designated people passing by predefined points of interest. Better tracking may facilitate better counting when people appear at designated point, but in the integrated framework, the value of tracking rests in accomplishing the high-level task, not an explicit of measure of how far tracks of people deviate from ground truth.

Generally, it is not a simple matter to decide what parts of a system to evaluate in isolation and when. There are always tradeoffs, and piecewise versus integrated evaluation considerations need to be explicitly considered early in the construction of data sets and associated evaluation protocols. A good rule of thumb is that evaluation involving performance measures not meaningful to an end user of a deployed system may be of only secondary importance.

Another major challenge for evaluation in general and certainly in the context of video analytics is to transition from a diagnostics only approach toward one that is predictive. The distinction roughly speaking goes as follows. Diagnostic evaluations make explicit when a system failed and succeeded over a constrained test set. Those responsible for drawing conclusions about the overall performance of the system may report summaries statistics such as error rates and even focus attention on some of the failure cases as partial explanation for why a system has failed.

Predictive evaluation is much more valuable as well as much harder. Consider the following very important question asked by most anyone attempting to deploy a new technology: "given my application domain has the following characteristics, how well can I expect the technology to perform?" The essential difference is that a predictive evaluation requires the construction of some form of performance model that relates the main characteristics of a task to a predicted level of performance. Such an assessment is extremely challenging to make when many performance-impacting variables are at play simultaneously. For example, the density of cameras vs. the density of the movers in a scene and changes in lighting, weather, and other naturally-varying parameters.

## 10.7 Data to Feed Research and Development Going Forward

Many, often conflicting, constraints come to bear on the problem of supplying data to advance research and development for video analytics. For a start, an open research and development environments accelerate the pace of advancement for video analytics. Further, to the extent the data used in open research and development projects is representative of operational tasks, rapid advancement will translate into robust and accurate fielded technology. The converse is also true, when research and development efforts are limited with respect to available data, and consequently those efforts are working on less than appropriate data, subsequent technology deployment often falters.

It's important to understand that operational problems versus open research and development have opposing sensitivities. Operations must be efficient, both in terms of cost and human effort. In addition, operations wont typically collect data beyond a difficulty threshold already imposed by the competence of current algorithms. In contrast, data sets constructed to promote open research and development carefully push the envelope in terms

of problem difficulty. Further, when data is collected specifically to foster open research and development, meta-data collection, including ground truth, can be anticipated and carried out, albeit at additional expense. Finally, privacy concerns are and important aspect of video analytics. When data is drawn from fielded operations, protecting privacy within research can be difficult. In contrast, when data collections are architected, privacy concerns may be addressed and handled appropriately in advance. However, there are significant trade-offs in realism. And, often artificialities in architected data can tuned to by machine learning algorithms resulting both in overstating of performance and optimizing to artificial conditions resulting in poor transition of the technology to operations.

One path forward that resolves some of these conflicting constraints begins by clearly delineating core technologies and, when possible, separating the essential elements from the surrounding operational trappings which often complicate real-world installations. Next, with the core technologies identified, then it is best to invest in well-designed semi-engineered data collection operations. Such data collections are not free, but as an investment they offer perhaps the greatest return relative to advancing the technology. They offer a means of demonstrating clearly that advancement is being made and hasten technology transfer through the avoidance of missteps.

Going down one level of detail, it will help if the following concrete actions are taken. First, foster technology partnerships between members of the research community and those with operational systems in controlled settings where instrumented data collection maybe carried out. Second, plan up front for how to manage privacy concerns relative to data collection and dissemination. For example, consider data collection at sites where prior agreement for use of data to support research is a reasonable request likely to be granted. Third, at the planning stages account for ways that both open and sequestered data of similar nature may be acquired and used in both open and more controlled formal (i.e. sequestered) evaluations. Fourth, invest significant resources into open architectures that support research and development, including "blind R&D" methods on sequestered data, as well as technology transfer in a continuous fashion. To be successful, such software must be of value to both researchers and those evaluating technology. When that condition is met, the benefits accrued to both sets of users is dramatic and the pace of technology deployment is accelerated.

A specific example of how many of the needs just outlined might be addressed is provided by a program being initiated by NIST in collaboration with DHS to stimulate research and development on video analytics. This program, the Video Tracking Analytics for Physical Security (VTAPS) Program, is adopting a challenge problem approach in order to stimulate the maturation and transition of video sensor network analytic technologies and standards to catalyze the creation of advanced infrastructure monitoring and forensic applications. A cross-disciplinary/cross-sector technical stakeholder workshop was held in July 2013 at NIST with over NNN researchers and government representatives attending. There are ongoing interagency discussions proceeding to refine the VTAPS Program. Work on an evaluation structure has already begun. A progression of challenge tasks leading toward 4D reconstruction of an observed environment using uncalibrated cameras is underway. Appropriate data collection is a critical need at this stage in the VTAPS Program.

One working model of how data might best be collected is in connection with monitoring a large public space, e.g. an airport. Figure 1 provides a motivating illustration. Several

things should stand out when reviewing Figure 54. First, simply note the scale of the space and associated need to integrate observations from many sensors relative to a complex 3D environment. Second, the high level nature of many of the most obvious questions a person might want answered relative to a facility such as an airport. For example, Show me on the map where this person I am pointing to has gone in the past hour. To offer another example: Show me people who appear to be watching escalator 3. These questions, and many others, hint at the depth of underlying technology advancements required before video analytics are truly up to the challenges imposed by such a setting. These advancements include real-time and forensic analysis, geo-spatial mapping, automatic video sequencing, biometrics and person identification, activity and event alerting, gaze and behavior analysis, and a privacy-preserving information flow.

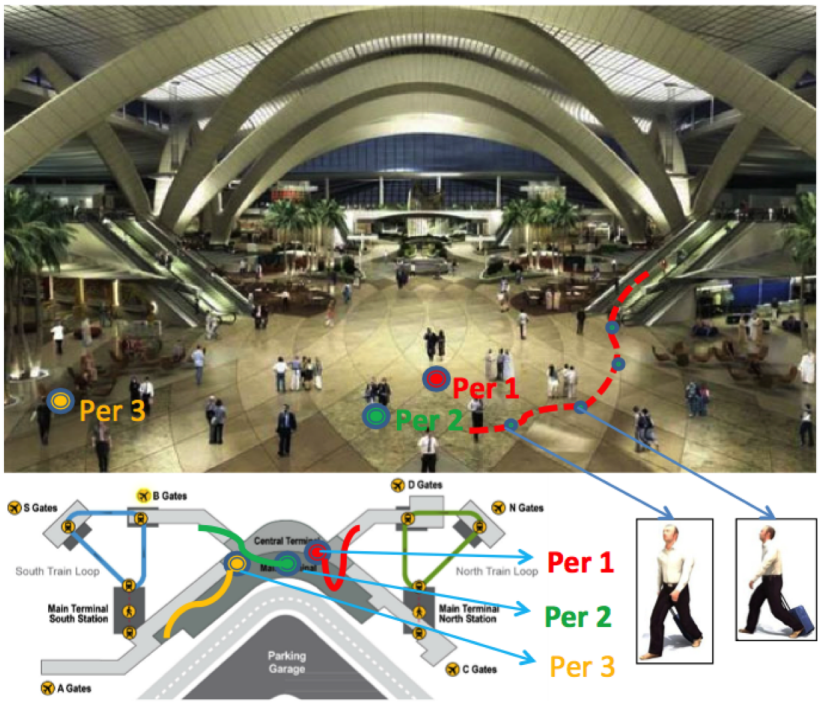



Figure 54: Concept Illustration for Airport Scenario.

## 10.8 A Case Study on Designing Challenges

Let us consider two options for an intended data collection.
Option 1 - One Billion People: A dataset with 2 images per person for 1 billion people. The people and images are sampled uniformly and independently from the larger population under a fixed scenario; e.g, Indias Unique Identification effort.

Option 2 - Two People: A dataset with1 billion images for each of two people. The images are provided from an oracle delivering images acquired under any set of desired circumstances.

It is likely that most readers of this report will quickly choose option number one. After all, imagery for one billion people represents a huge data set and appears a sound basis for technology advancement and experimentation. Indeed, there is much to argue for Option 1. However, the downside is too often overlooked. Option 1 is really only designed to answer a single question: "What is the performance on the data set with 2 images of 1 billion people collect in a specific scenario?

All subsequent questions such as: "How much better does an algorithm work in well lit environments?" or How does a person's appearance vary over time?" are more difficult to answer. Recall the statement that the data is randomly sampled for a fixed scenario. This random sampling implies the data is representative of a larger population of possible image pairs of people collected under a small set of conditions. However, just as important, it implies no information or control over additional factors.

Now consider Option 2. The one question that cannot be answered is that of absolute performance level over a large population of people. However, the great host of questions concerning personal variation in appearance can be answered in great depth, as can questions regarding any set of factors and circumstances of possible operational interest. One might even argue that Option 2 is preferable to Option 1 because it opens the door to measuring variation in appearance in ways Option 1 does not.

It is the panels view that a balance between the two extremes is desirable. It is outwardly attractive to focus on very large evaluations, where "large" is a simple function of how many raw images or videos are available. This attraction in turn has drawn the research and development community towards existing pools of operational/found data. For example, data sets constructed by scouring the World Wide Web for images. By analogy with Option 1 above, the found data strategy generates a great deal of data in a very short time. But also by analogy with Option 1, at best what can be said of the data is that it is sampled in a pseudorandom ad hoc fashion with little or no associated metadata.

In contrast, what might be called laboratory data is collected according to a plan drawn up in advance with ground truth either arising by construction and/or subsequent careful annotation. Given the human labor involved in collecting such data, in absolute terms such data sets are typically smaller than those scavenged from large private and public sources: in some sense more in keeping with Option 2. Also in keeping with Option 2, the associated metadata supports diagnostic analysis that actually reveals important properties of what is influencing the performance of the technology - going much beyond a simple statement of one algorithm achieves a 3% higher score than another. Last, but by no means least important, data collected with the intent of being distributed publicly can adhere to protocols that mean public release will be possible. Data scavenged from other sources carry a host of associated property right issues and frequently cannot actually be legally distributed.

## 10.9 Building a Good Challenge Problem

Taking to heart the lessons learned from the most successful past data sets and challenge problems, here are a few of the most important considerations when assembling a challenge problem. To start, the goals must be both grandiose and simple. An example of a good goal might be to increase verification rates for video based person recognition an order of magnitude in four years. A poorly expressed goal is by definition harder to state simply, but

symptoms of failure might include a goal statement conveyed through an eight to ten item bullet list with vaguely expressed objectives covering apparently unrelated functionality.

Next, when setting an explicit goal, it is important to ensure that based on everything currently known success will be within the grasp of the research and development community. The importance of this consideration is hard to overestimate, one of the ways to fail at launching a challenge problem that truly promotes technology development is to ask for such rapid advancement that the community fails to make meaningful progress and then the best researchers in the field walk away from the challenge problem.

Easy access to data and making data accessible to as wide a range of researchers as possible are both very important if a challenge problem is to gain traction and truly become a focus of advancement and attention by the research and development community. In this observation there is a word of caution to programs that value control over data at the expense of broader recognition of the task. To put things simply, once research groups start publishing papers on a data set, there is a natural drive for others to pick up and move the ball forward with the same data set. If those labs find their access to the data blocked, the value of that data and the associated advancement may stall.

The last major point is to not confuse data alone with what is typically here called a challenge problem. Data, absent associated infrastructure and well-defined protocols for experimentation, seldom leaves a lasting positive mark on the advancement technology. Not to put too fine a point on it, images and video absent anything else are not that hard to come by, and the real scientific value of that data only comes into play when there is a relatively complete infrastructure to support the challenge problem. This infrastructure includes clear experimental questions and associated metadata that allows replication of results. At a minimum, this infrastructure should typically include baseline algorithms and all associated software needed to run a baseline experiment. Once a research lab is provided with such an infrastructure as a starting point, it becomes straightforward for that lab to begin pushing the performance envelope and benchmark their progress relative to a known standard.

While choosing baseline algorithms for comparison, It is better to avoid using levels of performance associated with highly engineered commercial products to set the standard against with research is judged. Care needs to be taken to avoid the trap of cutting off basic research and true progress by prematurely killing interest in next generation technology. The primary purpose of baseline algorithms is to provide a common reference point against which to measure performance gains. It further helps when the internals of the baseline are well known and understood by the research and development community.

There is one last concern when it comes to building challenge problems. It is possible to point to very successful challenges and cite as one of their virtues the fact that they have been productively used to advance research for well over a decade. However, it is also possible to cite data sets that have long since ceased to be useful and yet continue to be picked up and re-examined by new generations of researchers. Perhaps the best constructive advice that can come from this observation is for those constructing challenge problems to consider carefully the following questions.

First, is there a well-defined progression of several subtasks that provide a ladder for researchers to climb: lower rungs for immediate progress and higher rungs for later years? Datasets for which this is true are more likely to be viewed favorably over time. Second, are important factors, for example indoor versus outdoor operation, or amateur versus pro-

fessional photographers, made explicit in the metadata. Data sets that are little more than large pools of undifferentiated data lend themselves to answering only a few questions typically of the form algorithm A does better than B. It harder for research on such data sets to remain valuable as time goes forward. Finally, does the clearly stated grandiose goal of the challenge related well to operational tasks? The weaknesses associated with data collected ad hoc from field operations have already been enumerated. That said, history indicates the best challenges are those with well constructed from non-operational data that nonetheless capture essential elements associated with a large set of operational tasks.

## 10.10 Conclusion

The maturity of an image or video analytics technology may be gauged by the data sets and associated protocols used to evaluate that technology. Operational data and found data often possess a desirable connection to the real-world. However, they generally lack metadata required for diagnostic evaluation. They also frequently come with impediments to open distribution. Data collection efforts designed to straddle the line between operational real-world considerations and the demands of researchers have a long tradition in Computer Vision. They are characterized by openly available data, clear associated metadata, well-defined evaluation protocols, and staged levels of difficulty to better track technology advancement. The time and effort required to construct such data sets and all that goes with them should be accounted when planning programs to advance image and video analytics.

# References


[1] B. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 3, no. 1, Feb 2007.

[2] Y. Li, T. Zhang, and D. Tretter, "An overview of video abstraction techniques," HP Laboratory, Tech. Rep. HPL-2001-191, July 2001.

[3] A. Money and H. Agius, "Video summarization: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, no. 2, Feb 2008.

[4] X. Zhu, A. Elmagarmid, X. Xue, L. Wu, and A. Catlin, "Insightvideo: toward hierarchical video content organization for efficient browsing, summarization and retrieval," *IEEE Transactions on Multimedia*, vol. 7, no. 4, Aug 2005.

[5] J. Oh, Q. Wen, S. Hwang, and J. Lee, "Video abstraction," in *Video data management and information retrieval*, S. Deb, Ed. Idea Group Inc. and IRM Press, 2004, pp. 321–346.

[6] C. Taskiran and E. Delp, "Video summarization," in *Digital Image Sequence Processing, Compression, and Analysis*, T. Reed, Ed. CRC Press, 2005, pp. 215–231.

[7] B. Shahraray and D. Gibbon, "Automatic generation of pictorial transcripts of video programs," in *Society of Photo-Optical Instrumentation Engineers*, 1995, pp. 512–518.

[8] S. Smoliar and H. Zhang, "Content-based video indexing and retrieval," *IEEE Multimedia Magazine*, vol. 1, no. 2, Summer 1994.

[9] H. Zhang, J. Wu, D. Zhong, and S. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognition*, vol. 30, no. 4, Apr 1997.

[10] A. Ferman and A. Tekalp, "Multiscale content extraction and representation for video indexing," in *Proceedings of Society of Photo-Optical Instrumentation Engineers*, 1997, pp. 23–31.

[11] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *International Conference on Image Processing*, 1998, pp. 866–870.

[12] A. Hanjalic and H. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 8, Dec 1999.

[13] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 123–130.

[14] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Unsupervised view and rate invariant clustering of video sequences," *Computer Vision and Image Understanding*, vol. 113, no. 3, Mar 2009.

[15] L. Xie, P. Xu, S. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden markov models," *Pattern Recognition Letters*, vol. 25, no. 7, May 2004.

[16] D. Zhong, R. Kumar, and S. Chang, "Real-time personalized sports video filtering and summarization," in *ACM International Conference on Multimedia*, 2001, pp. 623–625.

[17] N. Shroff, P. K. Turaga, and R. Chellappa, "Video prcis: Highlighting diverse aspects of videos," *IEEE Transactions on Multimedia*, vol. 12, no. 8, Dec 2010.

[18] M. Irani, P. Anandan, and S. Hsu, "Mosaic based representations of video sequences and their applications," in *IEEE International Conference on Computer Vision*, 1995, pp. 605–611.

[19] J. Wang and H. Adelson, "Representing moving images with layers," *IEEE Transactions on Image Processing*, vol. 3, no. 5, Sept 1994.

[20] N. Vasconcelos and A. Lippman, "A spatiotemporal motion model for video summarization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1998, pp. 361–366.

[21] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg, "Webcam synopsis: peeking around the world," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[22] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Transactions on Graphics*, vol. 27, no. 3, Aug 2008.

[23] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1346–1353.

[24] R. Anirudh and P. Turaga, "Interactively test-driving an object detector: Estimating performance on unlabeled data," in *IEEE Winter Conference on Computer Vision*, 2014.

[25] IARPA, "Finder program," http://www.iarpa.gov/Programs/ia/Finder/finder.html.

[26] DARPA, "Vmr program," http://www.darpa.mil/Our_Work/I2O/Programs/Visual_Media_Reasoning_(VMR).aspx.

[27] G. Baatz, O. Saurer, K. Kser, and M. Pollefeys, "Large scale visual geo-localization of images in mountainous terrain," in *European Conference on Computer Vision*, 2012, pp. 517–530.

[28] DARPA, "First (eccv 2012) and second (cvpr2013) international workshops on visual analysis and geo-localization of large-scale imagery," (http://vision.eecs.ucf.edu/eccv-2012-workshop/)(http://vision.eecs.ucf.edu/cvpr-2013-workshop/).

[29] J. Hays and A. A. Efros, "Im2gps: estimating geographic information from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[30] G. Vaca-Castano, A. R. Zamir, and M. Shah, "City scale geo-spatial trajectory estimation of a moving camera," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1186–1193.

[31] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3d point clouds," in *European Conference on Computer Vision*, 2012, pp. 15–29.

[32] M. Bansal, K. Daniilidis, and H. S. Sawhney, "Ultra-wide baseline facade matching for geo-localization," in *First International Workshop on Visual Analysis and Geo-Localization of Large-Scale Imagery held in conjunction with the European Conference on Computer Vision*, 2012, pp. 175–186.

[33] M. Bansal, H. S. Sawhney, H. Cheng, and K. Daniilidis, "Geo-localization of street views with aerial image databases," in *ACM Multimedia*, 2011, pp. 1125–1128.

[34] E. Kalogerakis, O. Vesselova, J. Hays, A. A. Efros, and A. Hertzmann, "Image sequence geolocation with human travel priors," in *International Conference on Computer Vision*, 2009, pp. 253–260.

[35] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, 2004.

[36] ——, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, 1999, pp. 1150–1157.

[37] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 2, 2008.

[38] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2137–2144.

[39] ——, "Recovering surface layout from an image," *International Journal of Computer Vision*, vol. 75, no. 1, Oct 2007.

[40] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," in *European Conference on Computer Vision*, 2010, pp. 352–365.

[41] ——, "Finding things: Image parsing with regions and per-exemplar detectors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3001–3008.

[42] C. Wah, S. Branson, P. Perona, and S. Belongie, "Multiclass recognition and part localization with humans in the loop," in *International Conference on Computer Vision*, 2011, pp. 2524–2531.

[43] L. Baboud, M. Cadik, E. Eisemann, and H.-P. Seidel, "Automatic photo-to-terrain alignment for the annotation of mountain pictures," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[44] G. Schindler, P. Krishnamurthy, R. Lublinerman, Y. Liu, and F. Dellaert, "Detecting and matching repeated patterns for automatic geo-tagging in urban environments," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[45] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[46] M. J. Cummins and P. M. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *International Journal on Robotics Research*, vol. 27, no. 6, 2008.

[47] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *European Conference on Computer Vision*, 2010, pp. 748–761.

[48] W. Zhang and J. Kosecka, "Image based localization in urban environments," in *3DPVT*, 2006.

[49] A. R. Zamir and M. Shah, "Accurate image localization based on google maps street view," in *European Conference on Computer Vision*, 2010, pp. 255–268.

[50] D. M. Chen, G. Baatz, and Others, "City-scale landmark identification on mobile devices," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 737–744.

[51] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2d-to-3d matching," in *International Conference on Computer Vision*, 2011, pp. 667–674.

[52] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2599–2606.

[53] E. Bork and J. Su, "Integrating lidar data and multispectral imagery for enhanced classification of rangeland vegetation: A meta-analysis," in *Remote Sensing of Environment Journal*, 2007.

[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1106–1114.

[55] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CoRR abs/1311.2524*, 2013.

[56] N. Jammalamadaka, A. Zisserman, M. Eichner, V. Ferrari, and C. V. Jawahar, "Has my algorithm succeeded? an evaluator for human pose estimators," in *European Conference on Computer Vision*, 2012, pp. 114–128.

[57] P. Matikainen, R. Sukthankar, and M. Hebert, "Model recommendation for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2256–2263.

[58] S. Kluckner, T. Mauthner, P. Roth, and H. Bischof, "Semantic classification in aerial imagery by integrating appearance and height information," in *Asian Conference on Computer Vision*, 2009.

[59] S. Kluckner and H. Bischof, "Large-scale aerial image interpretation using a redundant semantic classification," in *International Society for Photogrammetry and Remote Sensing, Photogrammetric Computer Vision and Image Analysis*, 2010.

[60] Gould, T. Gao, and D. Koller, "Region-based segmentation and object detection," in *Advances in Neural Information Processing Systems*, 2009.

[61] B. C. Matei, H. S. Sawhney, S. Samarasekera, J. Kim, and R. Kumar, "Building segmentation for densely built urban regions using aerial lidar data," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[62] V. Verma, R. Kumar, and S. C. Hsu, "3d building detection and modeling from aerial lidar data," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2213–2220.

[63] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, "What makes paris look like paris?" *ACM Transactions on Graph*, vol. 31, no. 4, 2012.

[64] Y. Ke, J. Quackbush, and J. Im, "Synergistic use of quickbird multispectral imagery and lidar data for object-based forest species classification," in *Remote Sensing of Environment Journal*, 2007.

[65] Y. Sheikh and M. Shah, "Trajectory association across multiple airborne cameras," *IEEE Transactions on Pattern Analysis And Machine Intellegence*, vol. 30, no. 2, Feb 2008.

[66] M. Trauring, "Automatic comparison of finger ridge patterns," *Nature*, vol. 197, 1963.

[67] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," *Journal of the Acoustic Society of America*, vol. 35, 1963.

[68] W. W. Bledsoe, "Man-machine facial recognition," Panoramic Research Inc, Tech. Rep. PRI 22, 1996.

[69] A. J. Mauceri, "Feasibility study of personal identication by signature verication," North American Aviation, Tech. Rep. SID65-24, 1965.

[70] R. H. Ernst, "Hand id system," Patent US 3 576 537, 1971.

[71] J. G. Daugman, "High condence visual recognition of persons by a test of statistical independence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, 1993.

[72] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, 2004.

[73] C. L. Wilson, "Fingerprint vendor technology evaluation 2003: Summary of results and analysis report," NIST, Tech. Rep. NISTIR 7123, June 2004.

[74] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts Amherst, Tech. Rep. 07-49, October 2007.

[75] Y. Taigman, M. R. Ming Yang, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[76] U. Uludag, A. Ross, and A. K. Jain, "Biometric template selection and update: A case study in fingerprints," *Pattern Recognition*, vol. 37, no. 7, 2004.

[77] R. Gopalan and D. Jacobs, "Comparing and combining lighting insensitive approaches for face recognition," *Computer Vision and Image Understanding*, vol. 114, no. 1, 2010.

[78] X. Zhang and G. Yongsheng, "Face recognition across pose: a review," *Pattern Recognition*, vol. 42, no. 11, 2009.

[79] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino, "Impact of articial gummy fingers on fingerprint systems," in *SPIE Optical Security and Counterfeit Deterrence Techniques IV*, 2002, pp. 275–289.

[80] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoong," in *IEEE BIOSIG*, 2012.

[81] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoong database with diverse attacks," in *IAPR International Conference on Biometrics*, 2012, pp. 26–31.

[82] I. Chingovska, A. Anjos, and S. Marcel, "Anti-spoofing in action: joint operation with a verification system," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 98–104.

[83] B. Biggio, Z. Akhtar, G. Fumera, G. L. Marcialis, and F. Roli, "Security evaluation of biometric authentication systems under real spoofing attacks," *IET Biometrics*, vol. 1, no. 1, 2012.

[84] S. Yoon, J. Feng, and A. K. Jain, "Altered fingerprints: Analysis and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, March 2012.

[85] K. Ricanek, "The next biometric challenge: Medical alterations," *IEEE Computer*, vol. 46, no. 9, 2013.

[86] W. J. Scheirer, W. Bishop, and T. E. Boult, "Beyond pki: The biocryptographic key infrastructure," in *Springer Security and Privacy in Biometrics*, 2013, pp. 45–68.

[87] A. K. Jain, A. Ross, and U. Uludag, "Biometric template security: Challenges and solutions," in *European Signal Processing Conference*, 2005, pp. 469–472.

[88] J. Breebaart, C. Busch, and a. E. K. J. Grave, "A reference architecture for biometric template protection based on pseudo identities," in *IEEE BIOSIG*, 2008, pp. 25–38.

[89] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *International Conference of Computer Vision*, 2009, pp. 365–372.

[90] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar, "Describable visual attributes for face verification and image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, 2011.

[91] B. Heflin, W. Scheirer, and T. E. Boult, "Detecting and classifying scars, marks, and tattoos found in the wild," in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2012, pp. 31–38.

[92] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2751–2758.

[93] K. Duan, D. Parikh, D. Crandall, and K. Grauman, "Discovering localized attributes for fine-grained recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3474–3481.

[94] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult, "Multi-attribute spaces: Calibration for attribute fusion and similarity search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2933–2940.

[95] http://www.allthingsdistributed.com/2007/02/help_find_jim_gray.html, note = , title = Help Find Jim Gray, author = W. Vogels.

[96] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1349–1380, Dec 2000.

[97] Y. Rubner, C. Tomasi, and L. Guibas, "The earth movers distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, 2000.

[98] N. Rasiwasia, P. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Transactions on Multimedia*, vol. 9, no. 5, 2007.

[99] T. Minka and R. Picard, "Interactive learning with a society of models," *Pattern Recognition*, vol. 30, no. 4, 1997.

[100] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, vol. 4, no. 2, 2011.

[101] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[102] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, 2012.

[103] http://www.qualcomm.com/solutions/augmented-reality, note = , title = Qualcomm Vuforia, author = .

[104] http://strata.oreilly.com/2011/09/crowdsourcing-science-twitter-google-senate.html, note = , title = Crowdsourcing and gaming spur a scientific breakthrough, author = A. Watters.

[105] http://blog.eyewire.org/about/, note = , title = A game to map the brain, author = Eyewire.

[106] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313.5786, 2007.

[107] http://www.image-net.org/, note = , title = ImageNet, author = ImageNet.

[108] A. Borji and L. Itti, "State-of-the-art ion visual attention modelling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[109] D. Gao and N. Vasconcelos, "Decision-theoretic saliency: computational principles, biological plausibility, implications for neurophysiology, and psychophysics," *Neural Computation*, vol. 21, 2009.

[110] A. Borji, D. Sihite, and L. Itti, "Salient object detection: a benchmark," in *European Conference in Computer Vision*, 2012.

[111] S. Lu and N. Vasconcelos, "Learning optimal seeds for diffusion-based salient object detection," in *IEEE Conference in Computer Vision and Pattern Recognition*, 2004.

[112] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, 2007.

[113] J. Costa-Pereira and N. Vasconcelos, "Cross-modal domain adaptation for text-based regularization of image semantics in image retrieval systems," *Computer Vision and Image Understanding*, 2014.

[114] L. Fei-Fei, R. Fergus, and P. Perona, "A bayesian approach to unsupervised one-shot learning of object categories," in *IEEE International Conference on Computer Vision*, 2003, pp. 1134–1141.

[115] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European Conference on Computer Vision*, 2010, pp. 213–226.

[116] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *IEEE International Conference on Computer Vision*, 2011, pp. 999–1006.

[117] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *IEEE Conference on Computer Vision on Pattern Recognition*, 2012, pp. 2066–2073.

[118] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, 2000.

[119] P. Woodland, "Multiple instance boosting for object detection," in *ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition*, 2001.

[120] M. Dixit, N. Rasiwasia, and N. Vasconcelos, "Adapted gaussian models for image classification," in *IEEE Conference on Computer Vision on Pattern Recognition*, 2011, pp. 937–943.

[121] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias," *Machine Learning*, vol. 28, 1997.

[122] R. Raina, A. Ng, and D. Koller, "Constructing informative priors using transfer learning," in *International Conference on Machine Learning*, 2006, pp. 713–720.

[123] C. Do and A. Ng, "Transfer learning for text classification," in *Advances in Neural Information Processing Systems*, 2005.

[124] J. Yang, R. Yan, and A. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *ACM International Conference on Multimedia*, 2007, pp. 188–197.

[125] A. Bergamo and L. Torresani, "Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach," in *Advances in Neural Information Processing Systems*, 2010, pp. 181–189.

[126] Y. Aytar and A. Zisserman, "Tabula rasa: Model transfer for object category detection," in *International Conference on Computer Vision*, 2011, pp. 2252–2259.

[127] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu, "Translated learning: Transfer learning across different feature spaces," in *Advances in Neural Information Processing Systems*, 2008, pp. 353–360.

[128] G. Qi, C. Aggarwal, and T. Huang, "Towards semantic knowledge propagation from text corpus to web images," in *ACM International Conference on World Wide Web*, 2011, pp. 297–306.

[129] A. Torralba and A. Efros, "Unbiased look at dataset bias," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[130] J. Smith, M. Naphade, and A. Natsev, "Multimedia semantic indexing using model vectors," in *IEEE International Conference on Multimedia and Expo*, 2003, pp. 445–448.

[131] N. Vasconcelos, "From pixels to semantic spaces: Advances in content-based image retrieval," *IEEE Computer*, vol. 40, no. 7, 2007.

[132] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *IEEE International Conference on Computer Vision on Pattern Recognition*, 2009, pp. 1778–1785.

[133] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *European Conference on Computer Vision*, 2010, pp. 776–789.

[134] Y. Li, N. Snavely, and D. P. Huttenlocher, "Location recognition using prioritized feature matching," in *European Conference on Computer Vision*, 2010, pp. 791–804.

[135] N. Rasiwasia and N. Vasconcelos, "Scene classification with low-dimensional semantic spaces and weak supervision," in *IEEE International Conference on Computer Vision on Pattern Recognition*, 2008.

[136] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, "Scene recognition on the semantic manifold," in *European Conference on Computer Vision*, 2012, pp. 359–372.

[137] T. Lee, C. Yang, R. Romero, and D. Mumford, "Neural activity in early visual cortex reflects behavioral experience and higher-order perceptual saliency," *Nature Neuroscience*, vol. 5, 2002.

[138] M. Carandini and D. Heeger, "Normalization as a canonical neural computation," in *Nature Reviews Neuroscience*, 2012.

[139] T. Dietterich, R. Lathrop, and T. Lozano-Perez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, 1997.

[140] O. Maron and A. Ratan, "Multiple instance learning for natural scene classification," in *International Conference in Machine Leaning*, 1998.

[141] Y. Chen and J. Wang, "Image categorization by learning and reasoning with regions," *Journal of Machine Learning Research*, 2004.

[142] P. Viola, J. Platt, and C. Zhang, "Multiple instance boosting for object detection," in *Advances in Neural Information Processing Systems*, 2005.

[143] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, 2009.

[144] R. Vezzani, D. Baltieri, and R. Cucchiara, "People re-identification in surveillance and forensics: a survey," *ACM Computing Surveys*, vol. 46, no. 2, 2014.

[145] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, 2014.

[146] S. Gong, M. Cristani, S. Yan, and C. Loy, "Person re-identification," in *Advances in Computer Vision and Pattern Recognition Series Springer*, 2014.

[147] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *British Machine Vision Conference*, 2011, pp. 68.1–68.11.

[148] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer Vision and Image Understanding*, vol. 117, no. 2, Nov 2013.

[149] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification : What features are important?" in *European Conference on Computer Vision Workshops and Demonstrations*, 2012, pp. 391–401.

[150] N. Martinel and C. Micheloni, "Re-identify people in wide area camera network," in *International Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 31–36.

[151] Y. Wu, M. Minoh, M. Mukunoki, W. Li, and S. Lao, "Collaborative sparse approximation for multiple-shot across-camera person re-identification," in *Advanced Video and Signal-Based Surveillance*, 2012, pp. 209–214.

[152] N. Gheissari, T. Sebastian, and R. Hartley, "Multiple person re-identification using part based spatio-temporal color appearance model," in *International Conference on Computer Vision Workshop on Visual Surveillance*, 2011, pp. 1721–1728.

[153] A. Bedagkar-Gala and S. K. Shah, "Part-based spatio-temporal model for multi-person re-identification," *Pattern Recognition Letters*, vol. 33, no. 14, 2012.

[154] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Boosted human re-identification using riemannian manifolds," *Image and Vision Computing*, vol. 30, no. 6, June 2012.

[155] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, 2013.

[156] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *International Conference on Computer Vision and Pattern Recognition*, 2013.

[157] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *British Machine Vision Conference*, 2012, pp. 57.1–57.11.

[158] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Asian conference on Computer Vision*, 2010, pp. 501–512.

[159] G. Zhang, Y. Wang, J. Kato, T. Marutani, and M. Kenji, "Local distance comparison for multiple-shot people re-identification," in *Asian conference on Computer Vision*, 2013, pp. 677–690.

[160] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *International Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2288–2295.

[161] M. Hirzer, P. M. Roth, K. Martin, and H. Bischof, "Relaxed pairwise learned metric for person reidentification," in *European Conference Computer Vision*, 2012, pp. 780–793.

[162] M. Hirzer, P. M. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *Advanced Video and Signal-Based Surveillance*, 2012, pp. 203–208.

[163] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Asian Conference on Computer Vision*, 2012, pp. 31–44.

[164] A. Mignon and F. Jurie, "Pcca : A new approach for distance learning from sparse pairwise constraints," in *International Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2666–2672.

[165] S. Pedagadi, J. Orwell, and S. Velastin, "Local fisher discriminant analysis for pedestrian reidentification," in *International Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3318–3325.

[166] L. An, M. Kafai, S. Yang, and B. Bhanu, "Reference-based person re identification," in *Advanced Video and Signal-Based Surveillance*, 2013.

[167] D. Figueira, L. Bazzani, H. Minh, M. Cristani, A. Bernardino, and V. Murino, "Semi-supervised multi-feature learning for person re-identification," in *International Conference on Advanced Video and Signal-based Surveillance*, 2013.

[168] H. Minh, L. Bazzani, and V. Murino, "A unifying framework for vector-valued manifold regularization and multi-view learning," in *International Conference on Machine Learning*, 2013, pp. 100–108.

[169] L. Yang and R. Jin, "Distance metric learning : A comprehensive survey," Michigan State University, Tech. Rep., 2006.

[170] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," in *ArXiv e-prints*, 2013.

[171] F. Porikli and M. Hill, "Inter-camera color calibration using cross-correlation model function," in *IEEE International Conference on Image Processing*, 2003, pp. 133–136.

[172] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera spacetime and appearance relationships for tracking across non-overlapping views," *Computer Vision and Image Understanding*, vol. 109, no. 2, Feb 2008.

[173] A. Gilbert and R. Bowden, "Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity," in *European Conference Computer Vision*, 2006.

[174] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions," in *British Machine Vision Conference*, 2008.

[175] A. Datta, L. M. Brown, R. Feris, and S. Pankanti, "Appearance modeling for person re-identification using weighted brightness transfer functions," in *International Conference on Pattern Recognition*, 2012, pp. 501–512.

[176] C. Siebler, B. Keni, and R. Stiefelhagen, "Adaptive color transformation for person reidentification in camera networks," in *International Conference on Distributed Smart Cameras*, 2010, pp. 199–205.

[177] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch, "Learning implicit transfer for person re-identification," in *European Conference on Computer Vision Workshops and Demonstrations*, 2012, pp. 381–390.

[178] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino, "Re-identification with rgb-d sensors," in *European Conference on Computer Vision Workshops and Demonstrations*, 2012, pp. 433–442.

[179] P. Salvagnini, L. Bazzani, M. Cristani, and V. Murino, "Person re-identification with a ptz camera: an introductory study," in *International Conference on Image Processing*, 2013.

[180] O. Javed and K. Shafique, "Appearance modeling for tracking in multiple non-overlapping cameras," in *International Conference on Computer Vision and Pattern Recognition*, 2005, pp. 26–33.

[181] N. Gheissari, T. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *International Conference on Computer Vision and Pattern Recognition*, 2006, pp. 1528–1535.

[182] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank, "Principal axis-based correspondence between multiple cameras for people tracking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, April 2006.

[183] X. Wang, G. Doretto, T. S. andJens Rittscher, and P. Tu, "Shape and appearance context modeling," in *International Conference on Computer Vision*, 2007, pp. 1–8.

[184] K. wen Chen, C. chuan Lai, and Y. ping Hung, "An adaptive learning method for target tracking across multiple cameras," in *International Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[185] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching under illumination change over time," in *ECCV Workshop on Multi- Camera and Multi-Modal Sensor Fusion Algorithms and Applications*, 2008, pp. 1–12.

[186] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European Conference on Computer Vision*, 2008, pp. 262–275.

[187] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *British Machine Vision Conference*, 2009, pp. 1–11.

[188] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," in *International Conference on Advanced Video and Signal Based Surveillance*, 2010, pp. 435–440.

[189] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *International Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2360–2367.

[190] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *British Machine Vision Conference*, 2010, pp. 21.1–21.11.

[191] W. Li and X. Wang, "Locally aligned feature transforms across views," in *International Conference on Computer Vision and Pattern Recognition*, 2013.

[192] W.-S. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, June 2013.

[193] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos, international journal for multimedia information retrieval," in *Asian conference on Computer Vision*, 2012.

[194] Y. Yang, I. Saleemi, and M. Shah, "Discovering motion primitives for unsupervised grouping and one-shote learning of human actions, gestures, and expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, July 2013.

[195] H. Wang, A. Klaser, C. Schmid, and C. Liu, "Dense trajectories and motion boundary descriptors for action recognition," in *International Journal of Computer*, 2013.

[196] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[197] Y. Tian, R. Sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[198] M. Jain, J. C. van Gemert, H. Jgou, P. Bouthemy, and C. G. M. Snoek, "Action localization by tubelets from motion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[199] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviours in crowd scenes using stability analysis for dynamical systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, Oct 2012.

[200] S. Khokhar, I. Saleemi, and M. Shah, "Statistical event representation and similarity invariant classification by kl divergence minimization," in *International Conference on Computer Vision*, 2011.

[201] ——, "Multi-agent event recognition by preservation of spatiotemporal relationships between probabilistic models," *Image and Vision Computing*, vol. 31, no. 9, Sept 2013.

[202] http://crcv-web.eecs.ucf.edu/ICCV13-Action-Workshop/, note = , title = Action Workshop, author = A. Watters.

[203] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, 2010.

[204] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[205] D. Hoiem, A. Efros, and M. Hebert, "Geometric context from a single image," in *IEEE International Conference on Computer Vision*, 2005.

[206] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Scene parsing with multiscale feature learning, purity trees, and optimal covers," in *International Conference on Machine Learning*, 2012.

[207] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," *International Journal of Computer Vision*, vol. 101, 2013.

[208] ——, "Finding things: Image parsing with regions and per-exemplar detectors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[209] J. Yang, B. Price, S. Cohen, and M. Yang, "Context driven scene parsing with attention to rare classes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[210] A. Saxena, M. Sun, and A. Ng, "Make3d: Learning 3-d scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[211] A. Gupta, A. Efros, and M. Hebert, "Blocks world revisited: Image understanding using qualitative geometry and mechanics," in *Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics*, 2010.

[212] T. Pollard and J. Mundy, "Change detection in a 3-d world," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[213] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes, "Layered object models for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[214] "Gigapixel," http://www.gigapixel.com.

[215] E. Swears, A. Hoogs, and K. Boyer, "Pyramid coding for functional scene element recognition in video scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[216] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[217] J. Deng, J. Krause, and L. Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[218] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *International Conference on Computer Vision*, 2013.

[219] F. Han and S. Zhu, "Bottom-up/top-down image parsing with attribute graph grammar," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, Jan 2009.

[220] Y. Lu, B. Yao, Y. Wang, and S. Zhu, "Reconfigurable templates for robust vehicle detection and classification," in *Workshop on Application of Computer Vision*, 2012.

[221] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[222] B. Rothrock, S. Park, and S. Zhu, "Integrating grammar and segmentation for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[223] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua, "Efficient boosted exemplar-based face detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[224] Y. B. Zhao and S. Zhu, "Scene parsing by integrating function, geometry and appearance models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

# A Workshop Agenda

Workshop on Frontiers in Image and Video Analysis
Sponsored by NSF, FBI, DARPA and UMIACS (UMD)
January 28-29, 2014
Agenda

January, 28th

7:30 a.m. Registration

8:10 a.m Plenary session
Goals of the workshop (Donlon, Bataille, VorderBruegge, Geertsen, and Chellappa)

8:30 a.m. Session 1
Video summarization, shot detection, and/or scene change detection (Peleg, Srivastava, Thornton, Turaga)

Session 2
Visual Analysis and Geo-Localization of Large-Scale Imagery (Crisman, Effros, Irvine, Sawhney)

10:00 - 10:30 Break

10:30 a.m. Session 3
Image-based Biometrics (Boult, Jain, Jacobs, Kriegman, Savvides)

10:30 a.m. Session 4
Human in the loop (Chang, Grauman, Karam, Khosla, Vasconcelos)

12:00 - 1:00 Lunch (SRI Demonstration, 12:30 -12:50)

1:00 p.m. Plenary session
Summary of Sessions 1-4 (Jacobs, Sawhney, Turaga, Vasconcelos)

2:00 p.m. Session 5
Re-identification of Persons and Vehicles in Videos and Images (Murino, Roy-Chowdhury, S. Shah, Yang)

2:00 p.m. Session 6
Human Activity Understanding (detection and recognition) in Video (Hoogs, Nevatia, M. Shah, Vidal)

3:30 -3:45 p.m. Break

3:45 p.m. Session 7
Semantic summarization (attribute-based scene tagging) (Ettinger, Hebert, Shi)

3:45 p.m. Session 8
Large scale visual recognition (Darrell, Hauptman, Hua, Li, Zhu)

5:15 p.m. Summary of sessions 5-8 (Ettinger, Roy-Chowdhury, Shah, Zhu)

6:00 p.m. Day 1 closes

Day 2 (Jan. 29th)

8:00 a.m. Recap of day 1 and goals for day 2 (Chellappa, Davis)

8:30 a.m. Plenary session
Datasets and performance evaluation for Research in Large-Scale Video and Image Content Analysis (Beveridge, Bowyer, Garofolo, Kasturi, Phillips)

9: 15 a.m. Discussion of solved (i.e., ready to deploy in specific operational scenarios), Davis

10:30 a.m. Break

10:45 a.m. Discussion of nearly solved (i.e., 1-to-3 years to deployment) Hebert

12:00 a.m. Lunch

1:00 p.m. Discussion of Over-the-Horizon problems (i.e., those requiring concerted effort over the next 3-5 years and beyond) Chellappa

2:30 p.m. Wrap up and writing assignments