Understanding Activity: Learning the Language of Action

Randal Nelson and Yiannis Aloimonos
Univ. of Rochester and Maryland

1.1 Overview

Understanding observed activity is an important problem, both from the standpoint of practical applications, and as a central issue in attempting to describe the phenomenon of intelligence. On the practical side, there are a large number of applications that would benefit from improved machine ability to analyze activity. The most prominent are various surveillance scenarios. The current emphasis on homeland security has brought this issue to the forefront, and resulted in considerable work on mostly low-level detection schemes. There are also applications in medical diagnosis and household assistants that, in the long run, may be even more important. In addition, there are numerous scientific projects, ranging from monitoring of weather conditions to observation of animal behavior that would be facilitated by automatic understanding of activity. From a scientific standpoint, understanding activity understanding is central to understanding intelligence. Analyzing what is happening in the environment, and acting on the results of that analysis is, to a large extent, what natural intelligent systems do, whether they are human or animal. Artificial intelligences, if we want them to work with people in the natural world, will need commensurate abilities. The importance of the problem has not gone unrecognized. There is a substantial body of work on various components of the problem, most especially on change detection, motion analysis, and tracking. More recently, in the context of surveillance applications, there have been some preliminary efforts to come up with a general ontology of human activity. These efforts have largely been top-down in the classic AI tradition, and, as with earlier analogous effort in areas such as object recognition and scene understanding, have seen limited practical application because of the difficulty in robustly extracting the putative primitives on which the top-down formalism is based. We propose a novel alternative approach, where understanding activity is centered on perception and the abstraction of compact representations from that perception. Specifically, a system receives raw sensory input, and must base its "understanding" on information that is actually extractable from these data streams. We will concentrate on video streams, but we will presume that auditory, tactile, or proprioceptive streams might be used as well. There has been significant recent progress in what has been loosely termed "image-based" object and action recognition. The relevant aspect of the image-based approach is that the primitives that are assembled to produce a percept of an object or an action are extracted from statistical analysis of the perceptual data. We see this feature learning and first-level recognition as representative of the sort of abstraction that is necessary for understanding at all levels. We think that the statistical feature abstraction processes, and the structural "grammars" that permit them to be assembled in space (for object recognition) and time (for action recognition), can be extended to 1), reduce the human input required and 2), generate a higher level of abstraction. The resulting "concept extraction" process will produce a compact, extensible representation that enables event-based organization and recall, predictive reasoning, and natural language communication in a system observing activity in natural environments. This process of repeated abstraction and organization will naturally induce a "symbolic" structure onto the observed world. This approach is in contrast to certain classical approaches where understanding is based on analysis of input that is already in symbolic/linguistic form. Challenging problems can be found in this approach; however we feel that the state-of-the-art in extracting symbolic descriptions from real-world data remains so primitive that little can be assumed about the information that might be available. In fact, the extraction of the "symbolic" description *is* the primary problem, and the focus should be on making constructive use of what *can* be extracted, rather than on artificially formal problems constructed about what we hope *might* be extractable.

1.2 Understanding

We have used the term "understanding" repeatedly in the above introduction, without defining it precisely. What does it mean for an intelligent agent to "understand" activity? This question borders on the philosophical, and given the current state of the art, could be debated endlessly. However, we feel there is some leverage to be gained by attempting to define in a functional sense, what it might mean for an agent to "understand" activity that it observes. Several possibilities come to mind.

1. The agent is remembering what it observes. Specifically, it is storing a compact, indexed representation that captures salient aspects of the experience, related and linked to previous experience. In other words, understanding consists of constructing an integrated, accessible episodic memory. Abstracted "symbolic" structure arises as an emergent phenomenon from the requirements of compactness and indexability. Important issues include: how is saliency determined?, how are the compact representations acquired? should old stored experience be modified or re-arranged on the basis of new structuring mechanisms?

2. The agent is attempting to predict what it will observe next and/or nearby (or even in the past, had it been looking, or if memory is re-examined.) Again, an agent will typically be interested only in certain salient aspects of the environment, which implies some co-existing analytic framework. However, interesting structures and processing

strategies tend to arise even if the only task is to predict the raw sensory stream. Understanding thus consists of the ability to predict/infer aspects of the world that are not directly observable. Abstracted "symbolic" structure arises as a necessary component of being able to predict within a compact framework. Issues include all those associated with memory plus how to deal with the continuous nature of time vis-a-vis the discrete nature of the artifacts that facilitate compact prediction mechanisms. Specifically, predicting everything (e.g.) one second in advance seems not optimally useful, but predicting everything at all times in the future seems highly redundant and wasteful of effort. This might force the adoption of an "event" economy - symbols are localized in time as well as space.

3. The agent is capable of performing actions that it sees, that is, it is observing the activities of another agent with which it can identify *empathically*. "Understanding" consists of accessing the resources and routines required to perform an imitation and readying them for activation. This makes available a lot of information that the agent has about the activity, its requirements, and its execution; and instantiates a certain complex state in the agent. A sufficiently robust and flexible agent might then make other use of this information than simply emulating what it observes. Abstracted "symbolic" structure reflects a modular organization of the agent's sensory-motor abilities.

4. The agent has a definable goal and is attempting to use what it observes to maximize the likelihood of successful execution of that goal. This is pretty classic, and has been the basis for approaches at every level of abstraction from conventional planning in the pre-defined symbolic domain, to reinforcement learning in the few-states domain, to adaptive visual servoing in the continuous domain. One of the natural language communication methodologies we have explored is the idea of using explicit communication goals for generating most likely effective utterances on the basis of empathic knowledge of shared perceptual abilities and common environment, fits into this framework. At one level, the understanding required in each of the above functional scenarios is different. However, all of them involve some form of "symbolic" abstraction of the sensory stream. It is on this common abstraction requirement that we think research would be most profitably concentrated, specifically on automatic sensory abstraction processes and representational frameworks that can subserve all of the afore-mentioned functions. In order to ensure that we remain tied to realizable strategies, we believe that any such research should be pursued in the context of a specific implementation of one or more of the above "understanding" scenarios. One of the the most accessible (and practically attractive) is the creation of a referenceable episodic memory. Such an artifact would represent an intelligent human augmentation system that will interact with users via a natural language oriented interface. Clearly, such assistive systems will require sensory abstractions that facilitate the interpretation and generation of natural language utterances, as well as ones that support the specific applications. It is our belief that these will be one and the same: specifically, that abstractions that emerge for the purposes of functional understanding activity, will support natural language dialogue concerning that same activity. This idea, that abstractions that arise for the purposes of non-linguistic functionality are an important foundation of linguistic abilities, is an important fundamental concept.

2 A Vision

As a simple example of the type of assistive interaction that activity understanding would enable, consider the following (hypothetical) dialogue fragments.

The first takes place in a home environment, where the machine has access to multiple active sensors covering several rooms or workspaces, and archives observations of what they have observed.

User: Do you know where my math book is?

System: I don't know about your math book. What does it look like?

User: It's blue and orange.

System: How big?

User: (Pointing) About as big as that red one on the desk.

System: This one? (showing an image with an object highlighted)

User: Yes.

System: When did you last have it?

User: Last night I think, maybe in the living room.

System: I didn't see you leave it there.

User: Um.. .

System: (interrupting and showing an image) I found this in the family room where you were sitting last night.

User: Oh good, that's it. Thanks.

In the above example, it turns out that the system does not actually know what a book is (more specifically, it has no visual recognition referent connected to that name). It could be taught, but learning about generic objects is time consuming for the human instructor, and the system manages to solve the problem using its memories of activities that it has observed, and generic perceptual abilities without requesting a complex interaction. The system could take note of the deficit, and request instruction at an appropriate time, especially if the term re-occurs in several dialogues. In the second example, the system has access to multiple cameras that monitor the public streets and areas around an office complex.

User: Can you inform me when the bus stops at 42nd street?

System: Yesterday the bus arrived at about 10. I need to use my camera then to watch for a taxi for Jim.

User: I have another camera I can plug in.

System: Good. Can I use it to get more data for George's traffic analysis too?

User: Sure.

Here we have an example of the machine collaborating for its "own" benefit (in its role as a proxy for others) in the face of limited resources. It also employs a "today will be like yesterday" predictive model in its reasoning.

The above examples are not far beyond what could be accomplished by integrating existing perceptual and natural language understanding technology. The reasoning seems sophisticated, but could actually be produced by simple "greedy" optimization procedures. The main perceptual elements are some activity understanding mechanisms that attend to people and vehicles along with a notion of the salience of events such as stopping and starting, and leaving or picking up objects. The system also uses some generic abilities to describe objects (e.g. by color) recognize some specific objects (people, a bus) and associate names to places.

The segmentation, recognition, and descriptive processes appearing in the above examples are actually feasible with current technology, at least in known contexts. Thus if we know we are watching a street, it is possible to craft routines to detect and sort out the people from the cars from the trucks and busses with reasonable accuracy.

There are three main challenges toward producing a system such as the one described above. The first challenge isnone of abstraction: how the salient attributes, objects, and events by which the memory is organized, are determined.

In current systems, these are all painstakingly programmed by hand. The system is given (e.g) recognition procedures/ models for people, cars, and busses; for activities/events such as walking, waiting, crossing, picking up; and location names for streets, buildings, rooms, etc. Such systems may work well, but are not robust, and modification must usually be done at a low level by a system programmer. A more general system would be one where saliency could be determined through low intensity interaction with the user. This would require that the system do much of the attribute discovery and organization on its own, and only request user confirmation at a very high level. e.g. "I keep seeing this object. Is it important? What is its name?" "I see a lot of this stuff. Is there a name for it?" "This object occasionally stops here. Is that significant?" In other words, the system must be responsible for most of its own abstraction.

The second challenge is the generation and maintenance of the episodic memory that supports the described interactions: what is stored in the memory and how is it indexed and organized so as to facilitate the rapid search and retrieval needed to support the sort of interactions described. As a start, it is clear that the memory must be indexed both by "symbolic" activity and object attributes, and by less name-like "looks like this" descriptors. In other words, the memory must be organized by elements at various levels in the abstraction hierarchy. There are significant questions about what should be stored in the memory, at what level of detail, how and by what features it is indexed, and to what degree it undergoes dynamic reorganization as additional observations are made and new ways of organizing and interpreting information are discovered.

The third challenge is using the abstracted and remembered information to perform useful operations. A primary application is supporting a dialogue with a human user about what is happening in the world, what has happened in the past, and what to look for in the future. This dialogue should occur using a natural language interface. We do not advocate that all the complexities of unrestricted language should be addressed at the outset; however it is necessary to robustly ground the semantics of some subset of language in perception in a way that matches the expectation of the human users. This aspect of natural language understanding has seen limited work, primarily because, until very recently, machine perception has not been capable of providing any robust abstractions to which natural language semantics could be attached.

2.1 Prime Directives

We have described a vision involving a robust, autonomous system that organizes sensory experience into useful knowledge, and uses this knowledge to perform useful operations. This general goal has inspired a lot of work in AI. The novel aspect of the approach we advocate is the degree to which the knowledge acquisition and organization is a plastic, automatic aspect of the system, amenable to low-bandwidth training by human collaborators rather than hardwired by the system designer. Ultimately, we would like the system to start with a minimum of hard-wired knowledge and functionality, and develop specific functionality through experience. This opens the question of what that minimum needs to be. We can imagine a system receiving some sensory stream, and (with possible human assistance) attempting to organize that stream into some meaningful structure. However, given an agent with any ability to move or interact with the world in a way that modifies the input it receives (a requirement for intelligent behavior,) the content of the sensory stream is dictated by the system behavior. There must be some fundamental behavioral principles for a system to get off the ground.

We propose the following three "prime directives" as a basic framework. Although these are certainly not sufficient, they provide a useful vehicle for recognizing and structuring the sometimes very large amount of implicit knowledge that is hidden in the algorithms, data structures, and mathematical formalisms employed in "learning" systems.

1. Survive. The system should not do anything that causes damage to itself. In some systems, much or all of this may be hardwired, or a consequence of physical engineering. In addition, some low-level reflexes, such as balancing upright for a bipedal robot, might be best placed under this heading. In mobile systems, this directive clearly interacts extensively with the plastic aspects of the system. In primarily observational systems (e.g. those consisting of pan-tilt-zoom cameras,) there seems to be much less physical activity under program control that would allow the systems

to damage themselves (though, we did find it necessary in one of our experimental systems to program in a reflex to prevent the system from staring at the sun for long periods with the iris wide open.) However, it turns out that the survival directive can be expanded to protecting useful information and abstractions that the system has acquired. For example, in an experimental system that adaptively learned about day and night by fitting a dynamic, bimodal model to brightness measurements, we needed to protect against an extended period of bright or dark (e.g. the system getting shut in a closed room for a week) destroying the model. This aspect of survival is likely to be a significant issue as we start generating valuable abstractions and interpretations.

2. Be Responsible. In general, a system will have a set of what can be loosely termed "goals" or responsibilities that it needs to accomplish. These represent its purpose in life. A system might start with some hardwired responsibilities but in general, we want to investigate the degree to which this aspect can be left plastic, and subject to modification by a human coach. This is where saliency arises.

3. Be Curious. From the standpoint of aquiring abstractions, this is the most interesting of the directives. The idea is that absent any explicit responsibility or survival constraints, the system should make use of its interactive ability to seek out novel sensory experience and attempt to organize it. The results of this process will be tentative methods of organizing and abstracting the sensory experience that can be presented to a human user for evaluation and suggestions for refinement. This requires that certain basic abstraction capabilities be built in to the system. The interesting question is what these abilities should be, and how general can they be made. For instance, does the system need to be given a specific capability to learn to recognize 3D objects, or is this ability a consequence of a more general ability? Does it need to be given the notion that an object can be described by a color or is this emergent from some more general ability? Does it even need to start with the idea of an object or motion or are even these emergent from some basic information principle? (The answer to this last question might be yes, objects are emergent, but so universal that there is no point in taking the time to learn it for an individual system.)

3 Prior Work

The problems of abstraction have relevance not only for activity understanding, but the study of general cognition and human behavior. We believe, however, that robust understanding of activity encompasses many of the central issues of the general problem, and provides a singularly useful focus for grounding the important issues. Various problems associated with automated action recognition from video sequences have seen considerable attention in the computer vision community, with limited success. We feel that many of these are engineering problems that can be addressed successfully only when we have attained a better grasp of the underlying abstraction processes and representations. Broadly speaking, there exist two hypotheses that may explain the representations and processes of visual action understanding. The first, referred to as the "logical hypothesis," states that action understanding is based on formal analysis of the different logical elements that constitute the action. For example, when we observe a hand grasping a pencil, the analyzed elements would be the hand, the pencil, and the movement of the hand toward the pencil. According to this hypothesis, the association of these elements and inferences about their interaction suffice to provide the observer an understanding of the observed action. We should note that human action recognition from video has been actively studied in recent years, in part due to a variety of important applications in human/machine interaction and in surveillance, and that much of the existing work falls under the logical hypothesis. These techniques focus on detection and tracking of body parts; predicates describe individual motions (e.g., "RAISE(leg)"); and actions are treated as collections of predicates over time, and a number of techniques have attempted to model these in various ways. For example, Pinhanez and Bobick [39] assumed that actions can be decomposed into sub-actions, some of which could be detected directly by perceptual methods. They then exploit a variety of temporal constraints between occurrences of sub-actions using the framework of temporal logic developed by J. Allen. Such techniques have a variety of problems, however. One is that the associated computational problems are frequently NP-hard. In an extension, Ivanov and Bobick [21] employed stochastic parsing and used an extended parser to handle concurrent events. A drawback of this is that the parser needs to learn the probabilities of the productions. Pentland and his colleagues [49] used coupled HMM's and syntactic training data for this problem; this cannot easily handle concurrent sub-actions, however. Hongeng and Nevatia [18] converted Allen's temporal logic framework to a Bayesian network, which is rather computationally intensive and requires the evaluation of all possible combinations of sub-actions. A more basic problem, however is that these approaches finesse the fundamental abstraction and segmentation problems that appear in the real-world action domain. They also fail to take into account that an acting agent is a dynamical system that interacts with the physical world in space and time.

An alternative hypothesis, which we will refer to as the "visuomotor hypothesis," holds that we understand actions when we map the visual representation of the observed action onto our existing experience of reality, including prior observation and motor representation of the same action. The basic idea is that observation of an activity evokes cognitive processes that have been associated with similar observations in the past. This would include actions, objects, and contexts that were previously observed in conjunction with the triggering pattern. It also includes intentional state and motor processes associated with the the agent performing the action itself, if the activity is one the agent is capable of performing. This is consistent with recent

neurophysiological results involving “mirror neurons”. In fact, not only emulative motor programs, but responsive ones could be associated. If you see a ball thrown, you not only access the throwing routines, but the catching ones. Such motor associations emphasize the need to consider the dynamical system aspect of active agents. There is a considerable amount of work in object recognition, image databases, and video surveillance that is relevant to the representations and abstraction processes needed to understand activity in the functional sense. Some relevant aspects of this prior work are discussed below.

3.1 Image Databases

There has been a large amount of recent work and interest involving image and video databases, with the primary goal of providing efficient user access to the imagery contained therein, without expending excessive effort on human annotation. The work is relevant because of the relationship between the automatic organization and indexing employed by such systems and the abstraction procedures we seek. A number of systems have been developed having useful characteristics for some modes of access. Prominent early examples include IBM’s QBIC project [12, 14], the Photobook project at MIT [37], the Chabot Project at Berkeley [35], and ImageRover, [42] which finds pictures on the World Wide Web. There are even startup companies such as VIRAGE that offer software products and services for image database management.

The basic research issue has been to automatically organize the imagery in such databases by content that is meaningful to a human user. To date, indexing methods have generally used color, texture, and rough shape and geometry [5, 44, 6, 2]. Such properties can be used directly in “find me more pictures like this” mode, and have also been employed to take user “sketches” as inputs [24, 22]. A few systems have attempted to interact with the user to develop a model of some user concept in terms of available features. These are typically constraint systems rather than the generative abstraction processes we are interested in [28]. What has not been employed much in these systems is traditional object recognition and higher level abstractions of the sort that can provide semantic descriptions similar to those that could be provided by a human annotator (picture contains trains, automobiles, tables, chairs, dogs, cats, children playing, etc.) The main exception is face detection, which has a large amount of attention as a specific problem, and there are systems that can find faces in pictures fairly reliably e.g., [47]. There have been a few attempts to employ statistical correlation of low-level index features, such colored blobs to locate particular generic objects - mainly people and animals; some have met with modest success [15, 16]. More recently, the technique has been extended to words used to annotate art images and faces in the news. [1, 3]. The approach is interesting, but the results do not seem interpretable as the generation of abstractions corresponding to the words in question.

The idea of performing high-level abstraction on image or video databases in a minimally supervised manner has been little addressed until very recently. process [11]. One example, using real imagery, is the work by Picard to propagate texture labels through a database [38]. Recent work by Perona and others [13] in the training of recognition models from imagery labeled only as containing an object or not, has stimulated interest in this issue. Although limited in discrimination and coverage compared to the best recognition systems relying on supervised training, this work suggests that the sort of automated abstraction we are after might be possible.

3.2 Image-Based Object Recognition

Image-based object recognition methods were developed to make recognition systems more general than could be achieved with geometric models, and more easily trainable from visual data. These systems are of interest to us because their models are derived from imagery - a necessary quality of the abstraction processes we require. Much of the reported work is on static object recognition, but the same sort of techniques have been applied to action and other modes of visual perception. Most of these techniques essentially operate by comparing an image-like representation of object appearance against many prototype representations stored in a memory, and finding the closest match. They have the advantage of being fairly general in terms of the object types they can handle, and of being easily trainable, at least in a supervised sense.

The best known work in this area is the subspace methods which were applied first to the recognition of faces [46], and then generalized by Murase and Nayar to more general domains [30]. These methods first find the major principal components of an image dataset. The variation in appearance of an object is represented as a manifold in the subspace spanned by these components, and recognition is carried out by projecting images of unknown objects into the subspace and finding the nearest manifold. A large number of variations of the subspace technique exist. In order to deal with clutter and occlusion, which cause problems for subspace methods, a number of image-based techniques based on more local features have been developed. Huang and Camps [19] have adapted the subspace approach to segmented regions, thus obtaining some tolerance to clutter and occlusion. Viola and Wells [48] use a matching method based on mutual information, avoiding some of the problems with correlation. Schmid and Mohr [41] have reported good results for an appearance based system using local features based on differential invariants at interest points. Lowe [27], has developed a well known system that achieves recognition of 3D objects using scaleinvariant local features. Nelson and Selinger [33, 34] have demonstrated a technique based on contours in local context regions that handles 6 orthographic freedoms in the presence of clutter and modest occlusion. All the above-mentioned techniques use supervised approaches for learning the object representations from imagery, and require extensive segmented and labeled imagery. As the

number of objects recognized becomes larger, the process of collecting labeled and segmented training data becomes increasingly laborious. The problem of assembling representations from an un-labeled, and/or unsegmented corpus is harder. The most successful work to date has been using the sort of relevant feature-learning approach described, e.g., by Perona [13]. This work is promising, but can not yet achieve the discrimination and overall robustness to multiple geometric transformations, clutter, and occlusion displayed by the best supervised systems. Overall, the time seems ripe to push the investigation of minimally supervised training techniques for all aspects of visual perception, and extending techniques beyond rigid objects to activity and situational abstractions.

3.3 Video Surveillance

An area that has a lot in common with the perceptual aspects of activity understanding is video surveillance. Although most of this work is not specifically aimed at providing a foundation for abstraction and language attachment, it is a rich source of tracking algorithms and other perceptual processes that correspond to activity and event percepts.

The MIT Media Lab's Pfinder system [50] tracks colored blobs representing a user's head, hands, torso, and feet. Rehg et al. use motion, color and stereo triangulation to find users in front of a smart kiosk [40]. CMU and Sarnoff, collaborating in the VSAM project [26, 23, 7], detect and track multiple people and vehicles in cluttered scenes across a wide area using multiple active cameras. Simple human/vehicle/clutter classification is performed. Systems that detect simple activities involving interaction of tracked objects include Morris and Hogg [29] who statistically model interactions between people and cars. Maryland's W4 system [17, 9] detects and tracks multiple people, supplying approximate locations of major body parts, and detects carried objects and exchanges. The MIT AI Lab's forest of sensors project [45, 25] uses an adaptive background model to track moving objects, automatically calibrates several cameras with overlapping fields of view, classifies tracked objects and learns typical patterns of activity for a site so that unusual activities may be detected. The TI system [8] tracks objects and produces a graph representation of their spatio-temporal relations. A rule-based system detects different events involving the objects (appearance, disappearance, entrance, exit, deposit, removal, motion, and rest). Another approach to representing tracked movement in a scene is described by Irani and Anandan [20]. A method classifying more detailed human movement patterns is described by Davis and Bobick [10] who describe a view-based "temporal template" approach. Brand et al. [4] use coupled HMMs to represent gestures that involve both arms where the arms are independent but not completely decoupled. Oliver et al. [36] use the same framework to model human interactions upon meeting. They model five behaviors that capture different ways that people meet and either walk on together or go separate ways.

4 Foundations

The ultimate goal is to generate and organize abstract representations of activity with a minimum of human interaction. However, it is not necessary to start from scratch. We have previously developed robust, trainable systems for action and object recognition that can provide a starting point for deriving more general abstraction schemes, and for developing an episodic memory system. We have also also a preliminary framework for generating natural language utterances describing observed properties of natural environments.

4.1 View Invariant Recognition of Actions

See [51].

4.2 View Invariant 3D Object Recognition

Over the last few years, Rochester has developed a robust 3D object recognition system. In its original form, the system was trained from labeled imagery. Subsequent NSF-funded work demonstrated that the system could be modified to learn models from minimally annotated image collections. [34, 33, 43]. This work exploited the ability of the system to extrapolate to new poses of known objects, and to generalize to similarly shaped, but previously unencountered objects. The recognition system is based on a novel hierarchy of perceptual grouping processes. A 3D object is represented as a fourth-level perceptual group, consisting of a topologically structured set of flexible, 2-D views each derived from a training image. In these views, which represent third-level perceptual groups, the visual appearance of an object is represented as a geometrically consistent cluster of several overlapping local context regions. These local context regions represent high-information second-level perceptual groups, and are essentially windows centered on and normalized by key first-level features that contain a representation of all first-level features that intersect the window. The design and use of these high-information intermediate groups was the main innovation of our technique.

The first level features are the result of first level grouping processes run on the image, typically representing connected contour fragments, or locally homogeneous regions. The method has some important advantages. First, recognition is with full orthographic invariance, (two translational and three rotational degrees of freedom, plus scale). Second, because it is based on a merged percept of local features rather than global properties, the method is robust to occlusion and background clutter, and does not require prior object-level segmentation. This is an advantage over standard systems based on principal components and other forms of template analysis, which are sensitive to occlusion and clutter. Third, entry of objects into the memory can be an active, automatic procedure, where the system accumulates information from different views until sufficient information is acquired.

For the purposes of this project, the system gives us a robust starting ability to attach labels to objects encountered

during observation of the world. This capability has been demonstrated in a prototype memory assistance project, where a set of active pan-tilt-zoom cameras observed an apartment, and kept track of the location of a small set of important objects (portable phone, glasses, coffee cup) as the inhabitants moved them and other objects about the apartment [31]. The system incorporated a touch-screen, audio-visual interface where the user could query the location of an object by touching its picture, and receiving a highlighted image showing its location, and machine generated audio describing it, e.g., " your glasses are on the bedroom dresser". The system also stored a visual and symbolic history of the location of the objects of interest, as well as video clips showing the periods when they were picked up or set down. This can be seen as a preliminary version of the sort of episodic memory archive we advocate creating. The system also has an architecture within which automated abstraction mechanisms can be devised and evaluated. The current features and hierarchy were designed by hand. However, the only requirement on the features is that they have location, scale, and orientation associated with them. Thus automated abstraction mechanisms could be introduced at any level of the hierarchy, and the resulting performance compared to the hand-tailored system.

4.3 Grounded Utterances

Over the last two years we have made some preliminary efforts to robustly ground natural language in visual perception. This problem has proven quite difficult. We started with the goal of producing a general mid-level abstraction of visual perception that would support human interaction with a perceiving machine using natural language modalities.

One of our first discoveries was that we needed a model of the function of natural language interaction. We eventually formulated the issue as the problem of the speaker's inducing a desired change in the brain state or behavior of of the listener as efficiently as possible (given the low-bandwidth nature on NL communications). This led to a MLE (most likely effective) strategy that permitted the generation of a natural language utterance to be formulated as a statistical optimization problem employing empathic knowledge about the definition of visual concepts and their experienced statistical properties in a principled manner.

We were able to formulate abstractions for concepts such as named colors, and proportional adverbs (partly, mostly), and demonstrate effective description of generic objects automatically extracted from complex environments using this principle [32]. We also evaluated basic size and shape descriptors (large, small, thick, thin etc.) and basic relations (on, near, above etc.) and determined that the same strategy would apply. In the course of these investigations, however, it became clear that we were expending a large amount of knowledge engineering effort in the definition of the basic percepts on which the strategy operated. For example, we defined named colors as fuzzy functions over a normalized tri-chromatic space, and learned the parameters with a rather extensive supervised training procedure. The proportional adverbs were also fuzzy functions, but with some parametric assumptions of form, and over a one-dimensional space derived from integration over probabilistic component parts. Size/shape looked to be restricted low-dimensional functions over different subsets of a group of geometric and orientation measurements. The representation that allows robust recognition of objects, on the other hand, employs a representation consisting of associated sets of stored exemplars in an attentively defined, high dimensional space. An important realization ensued: although machine learning allowed us to (more or less) efficiently acquire models for individual examples of different concepts, a huge amount of work went into hand-tailoring the different conceptual classes and their representations (color, size, shape, object id etc). Hand tailoring enough of these to give broad competence seems difficult, and introduces a rigidity to the organization of knowledge. We need to make the process of concept development a more automatic process, which would be enabled by the automation of perceptual abstraction.

5 Research Challenges

This section describes some specific approaches we think might yield significant results in support of the goal of activity understanding.

5.1 Abstraction

What is abstraction? At the core, it involves a mechanism by which a (large) set of different instances are considered to be, in some sense, the same. In general the relation is considered to be, in some way, meaningful: i.e. there is a concise "explanation" for why the instances are grouped. An occurrence of one of these instances can thus, (in appropriate context) be replaced by a reference to the class. This permits both efficient representation of salient aspects of a situation, and efficient description of decision and control processes. Mapped to the world of patterns, an abstraction defines a set of patterns that belong together. The concise explanation condition is generally interpreted as requiring that the description of the set be (much) smaller than a simple enumeration of its members. A classic example of an abstraction in this sense would be the concept of a visual car with a pattern set (loosely) defined as the set of all (image, mask) pairs that represent a car in a picture. (We ignore for the moment issues of exactly what objects are cars, whether non-photographic drawings count, what if it is only one pixel, etc.) The point is, the set seems reasonably meaningful, and is clearly too large to be enumerated in any remotely practical setting.

It is hard to imagine the sort of abstraction needed to organize perceptual input in a manner convenient for human interaction occurring in a complete vacuum. The goal is to develop the sophisticated representations and saliency measures needed, with a minimum of user interaction. The prime directives of survival, responsibility, and curiosity provide some guidance. However, it is likely that some additional structuring elements will be necessary. The

processes might start by a user giving some examples of situations that were salient. These could be used as a starting point for extensions, which through a low bandwidth, interactive pruning process could result in the growth of a complex, ideosyncratic interpretation system. Even this requires some prior structure to get off the ground. We propose that this can be provided by some built-in (possibly implicit) knowledge of geometry and physics, fundamental notions of location, motion and objecthood, and some low level perceptual descriptors e.g for shape, marking, color, and texture. As described above, we have developed object and action recognition systems. Both of these go through a phase where they learn primitive features that are useful for modeling different objects/events. As the number of known objects increases, there begins to be some duplication in the primitives, and eventually a point is reached where few new primitives are needed to learn a new object/event. At this point learning at this level declines. Learning of new objects continues, however, and learning of structures in which these objects form primitives (eating a meal, reading a book) becomes possible. In fact, prior to the object/action feature stage, there are typically more primitive levels where very local features such as edges or boundaries or motion vectors are obtained. In some approaches these abilities are also learned or at least tuned up. The process of abstraction thus seems to be hierarchical. In current systems, such hierarchies are typically hand coded and fairly strict: e.g. pixels to edges, edges to contours, contours to salient groups, groups to views, views to objects, to scenes etc. Moreover, the learning or abstraction processes at each stage are typically different, and again, hand coded for the specific application. Although some hard-wired separation of levels may be necessary, we are interested in utilizing mechanisms that apply across levels, and that permit organization outside of a strict level hierarchy.

5.1.1 Approach

One approach to semi-automated abstraction would be to start with existing functionality developed in the areas of image classification, object recognition, and activity identification, and attempt to push two ways. First, attempt to increase the generality of the models, to allow them to describe a broader range of abstracted objects. Second, attempt to reduce the amount of human interaction that is needed to train the systems, primarily by devolving as much responsibility as possible onto the systems for identifying significant correlations in perceptual data and proposing abstractions based on them. The three prime directives will be used as a guide.

Image Level The lowest level concerns properties of images, or local sections thereof. At this level people (and animals) seem to formulate abstractions concerning environments, places, times, and materials. We could start with some pre-defined measures, e.g. global and regional means and variation of intensity, gradients, texture and color measures. We can hand-code some abstractions based on these in order to jump-start the episodic memory; however our main interest will be to put data-mining tools to work as these image statistics accumulate over time. A simple example is to use EM + Gaussian mixture models, and look for statistically significant clustering into small numbers of components. The system will then interactively ask the user if proposed clusters are meaningful. For those that are, the system may request shared symbolic referents (i.e. words). This should discover concepts such as day and night, image classes distinguished by texture, material classes that occur frequently (pavement, sky, vegetation etc.) If users respond that a regularity is not meaningful, the system can factor it out and re-examine the residual. Minimum description length (MDL) strategies might be useful in evaluating the quality of clustering. As users confirm significant discoveries, the MDL measure can be modified so that value of description is weighted toward preserving significant features rather than just raw information. At this level, the curiosity directive just attempts to gather lots of different images. It is only loosely specified, as the available measures of novelty are primitive. Survival will be mainly concerned with preventing the destruction of significant classes if they are left somewhat plastic to accommodate environmental variation. There is no explicit responsibility.

Objects and Actions Level

Here we could start with a pre-wired ability to segment and track moving, compact objects. There is some possibility we could “evolve” an ability to segment novel static objects using “is this something” questions, but this would likely be a heavy load on the user. We could then attempt to cluster spatial and temporal descriptions of these tracked objects, starting with iconic descriptors (e.g. maps of intensity, edges, local motion). The foundation step would be to use MDL to search for good local features for reproduction. Once useful features are discovered, they can be pushed back through an MDL process to emphasize local features that preserve clusters that users verify as meaningful. It would be interesting to see whether concepts corresponding to traditional abstractions such as object color or general size/shape arise from such analyses.

At this point, a mechanism for discovery of abstraction schemas (described in the next section) could be deployed to discover geometric and other part-shuffling invariances. These and similar techniques have the potential to produce quite general and robust abstractions at the object and action level. We have previously developed significant capability in trainable methods for recognizing objects and actions. These will form the basis for an evaluation of the quality of the more autonomously discovered abstractions. They also provide a basis for jump starting the episodic memory system.

The system at this level has a pre-wired responsibility of attending to motion and compact “objects”. Curiosity is employed during this phase to expand the knowledge base of objects and actions by attending to activity and objects that do not fit existing abstractions. Survival is again, mostly concerned with preventing destruction or excessive drift

of abstractions.

Activities and Scenarios Level

At the next level we might applyMDL and datamining techniques to the abstractions produced in objects/actions and image/place levels to find statistically significant correlations. At this point, MDL may be less useful than more sophisticated datamining approaches, as significant events may not account for much of the "energy" even at the abstracted symbolic level. A human coach would likely be required to answer questions about whether events are significant, and to ask for names or NL descriptions.

We might anticipate discoveries reflecting regularities such as cars are generally on pavement and the bus often stops at a particular place. Such regularities could be represented using prototypes that include conserved components. At this level discovered abstractions start to looks like classic "knowledge" rules, except the "symbolic" aspect is incidental - a useful tag, representing a statistical regularity of perceptual data. Unlike the previous two levels, there is no significant body of generic hard-wired perceptual techniques with which to compare our results or jump-start the episodic memory. There is a lot of ungrounded "symbolic AI" work that assumes purportedly analogous knowledge as given, but perceptually, it is largely uncharted territory. At this level, we could incorporate responsibilities from the user (e.g. watch for the bus in evening). Curiosity is still important at this level, and it would be much more directed since we are looking for novel combinations that recur. This directive may provide a way out of the NP-completeness issues that plague formal statements of many knowledge mining problems. Survival necessitates freezing of lower-level representations. We can push individual concepts around, and even select for features back across one level, but once low-level features are an essential component of a large body of higher-level representations and memory, changing them conflicts with survival - it would destroy the utility of hard-earned higher-level abstractions.

5.1.2 Abstraction Schemas

Consider for a moment, an underappreciated aspect of visual abstraction: translation within the image. In the field of object recognition, this is usually considered a trivial and solved problem. If the only variation we have to contend with is image translation, we can solve the problem either by lining up the centroid of a candidate object with that of a model, or scanning the model across all object locations. What is overlooked, is that the reason translation is so "easy" is because of the way we have chosen to represent images in linear machine memory, and because conventional computers perform symbolic address-shifting arithmetic as a primitive operation.

The problem is not really so easy. Considered as a bag of pixels, without the (x,y) coordinate system, the patterns representing different translations of the same 2D object are very different for translations larger than the typical size of

variation within the object. Given the variation in resolution across the retina, and the fact that symbolic arithmetic and linear addressing are not thought to be basic neural operations, the situation in the brain seems analogously complex. The real issue is the ability to efficiently learn compact representations of objects that translate. An abstraction e.g. of the letter "A" should allow us to recognize it anywhere in an image. A conventional learning algorithm might generate such an abstraction after being shown many examples at different locations. However, for the letter "B" the entire process would have to be repeated. The notion of a translation transformation however, provides a mechanism that is independent of the particular pattern. Rather than a simple abstraction, it is what we might term an *abstraction schema* The interesting question is, given the underlying complexity of translation on an irregular substrate such as the retina, if it is possible to "learn" such a schema efficiently in any meaningful way, and out of which possibilities is the selection being made? It turns out that the answer is yes, and that the selection is over what amounts to the class of all possible structural distortions that are parameterized by a "sufficiently small" index set (sufficiently small could be several thousands if you get a chance to see them all a few times.) The situation can be visualized as a "window", each location of which is mapped to some different location on the retina, depending on the setting of the index parameter. Such a schema can be represented by a set of tables that describe the transforming functional for each setting of the index parameter. The entries in the tables can be efficiently learned by fixing the index parameter, and showing a set of different patterns on the retina. By pruning the possible sources each time a new pattern is observed, the mapping can be constructed in O(log(N)) trials where N is the number of retinal pixels. Overall, the entire table can be learned in O(K log(N)) work, where K is the number of table entries, which is about as efficient as could be hoped for. Now we reach the crux of the matter. We are not particularly interested in learning translational mappings. The interesting question is, supposing the machinery is in place to do the sort of learning described above, *what else* is it possible to learn? The answer, as mentioned above, is that we can efficiently learn *any* class of structural permutation mappings (ones that scramble a set of elements) that are parameterized by a reasonably sized index set. This clearly includes rotation and scaling operations, and these could be learned separately and composed with translation to produce a system that could translate, rotate, and scale on demand.

But the class also includes activities. For example, if we take time as an index parameter, then walking is (largely) a structural distortion of some key frame, and so are a lot of other movements. If we know how one person walks, we know how a novel person walks. Also note that in our demonstration of learnability, there is no requirement that either the component mappings or the index sets to be continuous (the learning procedure could probably be made even more efficient if they were). A lot of language abstractions seem to involve rescrambling of components. If the rescrambling possibilities are indexed by a sufficiently

small set, then that too might fit into this framework. This class of abstraction schemas, and its generalizations seem potentially to be a very valuable avenue of investigation.

5.1.3 Discovering Humans

Central to visually understanding human actions is the ability to visually locate humans and determine their body configuration. In other words, we need abstractions (amounting to models) of humans that allow us to compactly reference and describe salient information in video imagery of people. This particular aspect of the world is so important that we think it worthwhile to direct some specific design effort in that direction. Our goal should be to make these abstractions independent of the physical characteristics of the human involved, which excludes approaches based on tracking templates of body parts. Alternative video features that we can exploit include motion (regardless of appearance, the motion of body parts is constrained in a particular fashion), and a qualitative description of shape.

To construct a body model from motion, we might use training video consisting of a single person performing some action from one view. We could compute optical flow for each frame of this video and in each frame group together patches with the same optical flow. These patches are then tracked across the video and more patches are added as different body parts first come into motion. In the end, we would obtain patches of image regions which moved more or less coherently during the action. These emergent elements may roughly correspond to the body parts, and can be viewed lower level abstractions obtained from an MDL-type process.

We can now form a graph with these patches as nodes and edges between every two pairs of nodes. Given two patches, we have time series data for their position and velocity vectors. We weight each edge of this graph with the degree of correlation between the two patches, which can be computed from the mutual information of the two series. Edges with low mutual information are discarded from the graph. This can also be viewed in an MDL framework as a process of discovering additional coherent structure. The model can be refined further by fitting a parametric function to the correlation between the two series. Since the constructed graph will vary across different views, we store the graphs obtained from a discrete set of views. The collection of these graphs forms our model of the human body. To find humans in a new video, we perform the same motion analysis in that video and match the resulting graph against the ones obtained from training data to obtain plausible hypothesis of body “part” location.

As described, the model generation process is a supervised one. However, since the model is generative rather than discriminative, it can be made the basis of a clustering process where recurring actions are identified in and abstracted from unlabeled video data.

Eventually the graph will be augmented with a qualitative description of body part functionality. Features in an image often correspond to features in the physical world which have some functional capacity. For example, the fact that the foot is supporting the body implies that the edges around the foot region should involve structures involving contact with the ground and vertical structures being supported from it. If image features can be independently associated with such functional descriptions, this information can improve the detection of component body parts by confirming or disregarding patches obtained from the motion analysis. Such functional information also plays a crucial role in further analysis of the action as the functional descriptions serve as the basis for higher level functional understanding of actions. The source of such information will initially be through user interaction. Ultimately, we expect such information to arise from the discovery of correlations with abstractions arising from other sensory modalities (e.g. haptic, tactile, and stereo information.)

5.1.4 Grounding the natural language of activity

In prliminary language grounding work, we justified hand tailoring of concepts/abstractions on the grounds that we were attaching words to existing computational and representational constructs that served non-linguistic functions (e.g. recognition, navigation, manipulation). This deflected the hand tailoring onto the need for certain functional abilities to be produced by whatever generated the system in the first place (evolution, NASA engineers...) For example, in the case of color, we argued that color-histogram-like representations were useful identifying objects. The named color representation, in combination with the proportional adverbs could then be proposed as a means of saving memory, while preserving much of capability. However, human color-recognition of objects seems more consistent with a color-histogram representation than with a named color one. The named color representation seems more designed to evoke an approximation of the color histogram into the mind of a communication partner over a low-bandwidth channel, and there is a considerable amount of processing needed to convert the underlying recognition model into the communication model. And the proportional adverbs are associated with more than color.

So given this, what we should be looking for is a uniform mechanism for extracting communication models (e.g. the base symbols that are assembled into an NL message) from the various abstracted representations. This is easiest if there is some underlying uniform form for the abstractions. We are not there yet, but we anticipate that the abstractions arising from the common use of MDL-like principles in different domains will provide fertile ground on which to test generic forms.

There is also the question of whether the communication representations are directly generated for communication, or attached to another functionality. We suggest that the need to store large amounts of episodic memory is an important precursor ability to natural language. The same sort of symbolic coding that supports low-bandwidth communication, also supports efficient storage of events. We think that exploring the proposition that language exploits mechanisms that developed for efficiently storing

large amounts of experience could be a very worthwhile enterprise.

5.2 Organization of Episodic Memory

A large-scale episodic memory employed in interactive, man-machine collaboration on analysis of real-world activity will involve large amounts of stored information that needs to be efficiently accessed by a number of different indexing schemes. This information includes not only the low-level data, but more abstract representations generated and utilized by perceptual modules that perform recognition and other operations, and semantic information required by higher-level processes. Moreover, this information is not static, but subject to periodic updating, reorganization, and re-interpretation. Abstracted models may also be augmented or otherwise modified as a system evolves; certainly higher-level semantic knowledge is a fluid commodity. This implies the need for an active "memory manager" at least to keep index mechanisms up to date, and possibly to handle certain kinds of modification. For implementation purposes, information for various systems modules may be stored and managed separately.

However, given the expectation that "tell me about the statuette you got in India" and "tell me about this" (holding it up) will produce similar responses, it seems clear that cross-modal access to information is an important process in natural communication. A uniform mechanism for representing and accessing widely disparate forms of information would be at least conceptually interesting, and possibly of practical benefit as well. One idea for doing this with perceptually derived information is described briefly below. At the base of the memory system, we imagine an analog to a web page—an "engram" containing a bunch of grouped information, possibly of disparate forms that is uniformly accessible via some address (a URL). For example, such an engram might be generated for a single object, or from a more complex observed episode (e.g., a user placing a notebook on a desk). At the lowest level it would contain the raw imagery and/or audio; above that, parameterized segmentations of actors and objects with generic size, color, and relational descriptions; then perscript-level descriptions of some attended objects and action. An engram might also contain pointers to other episodes considered relevant— narrative descriptions either generated during some user interaction, or prepared in anticipation of such interaction.

The web page analogy provides two important contrasts to a traditional record or C structure. The first is a web page's lack of a priori structure (though of course free-form record-based data structures can also be devised). The second is the implication of a dynamically maintained, content-based indexing mechanism, provided on the web by the various search engines.

Actual web pages contain text, various representations of graphics, video, audio, and links to (addresses of) other

pages. Current indexing mechanisms operate almost exclusively on the textual content and links, despite lots of interest and research on how to index the rest. "Engram pages" would contain signal attributes of percepts (raw, or slightly processed data), information indicating what sort of percept the data represents, higher-level representations, and (like actual pages) links to other pages. The indexing mechanism would operate both on the various data representations associated with the percepts, and on the URL analogs—the links—which can be viewed as playing the role of symbols. The overall strategy described above applies not just to visually derived information, but to information derived from acoustic and other sources as well.

6 Proposed Application: A HotWorld Archive

We suggest that a valuable result of research along the lines we have advocated would be the ability to construct a dynamic "History of theWorld" or (*HotWorld*) archive for recording, storage and man-machine collaborative analysis, of activity in the world. This archive would contain current and recorded low-level signals, progressively higherlevel (semantically meaningful) data abstracted from those recordings, and knowledge mined from the high-level data. Representations used in the higher-level portions of a HotWorld would be designed to support natural language interaction. Such a HotWorld archive requires a dynamic storage facility with the capacity to contain multiple terabytes of centrally organized and indexed information, acquired over time from real-world environments. To be generally useful, it should not cover just a single environment, but permit the representation of multiple environments, large and small, using various data sources and supporting different interaction modalities. As a base for confirmatory reference and reanalysis, (and possibly unlike human memory) much of the raw data provided by multiple video and audio sensors can be retained. There are of course limits to what can be accomplished even with today's inexpensive storage; however we estimate that capacity sufficient to store a year's worth of video, at DVD levels of compression (about 12 TB) is sufficient for the prototyping of non-trivial applications. Storing such raw data, however, would be the least interesting part of the system. To be useful for collaboration, the sensory information must be analyzed to produce more abstract representations. The first stage of this abstraction could be incorporated into embodied models of the sensory process. In human vision, every sensor has significant processing and control located close to the sensory element. In previous work we developed and implemented visual "computational sensors" in which a powerful workstation controls and processes data from one or two pan-tilt-zoom cameras. These sensors can be tasked with operations as complex as "watch for someone to place an object in the environment, zoom in and identify such objects if possible, and upload any objects discovered, including a parametric representation, their identity if known, and all raw data employed in the detection, representation and identification processes, along with linkages that allow the computations to be efficiently re-examined.

Though fairly powerful, the computational sensors are still limited in the abstractions they can perform in real time.

As data from multiple streams arrives at the storage facility, we will want to perform a variety of further analyses. These include the ongoing development of abstractions, correlation with other streams, and correlation, searches, and synthetic operations on large amounts of previously stored information. These computations will often be both irregular, especially relative to a particular sensor, and highly computationally demanding, suggesting the need for powerful clustered or shared-memory compute engines linked to the storage facility. Fortunately, such computational power is now available at reasonable cost, making such an archive, at least in terms of hardware, a plausible endeavor.

Bibliography

References

[1] K. Barnard, P. Duygulu, and D. A. Forsyth. Clustering art. In *Proc. Computer Vision and Pattern Recognition*

(CVPR01), pages 434–441, Kauai Hawaii, Dec 2001.

[2] J. Barros, J. French, W. Martin, P. Kelly, and M. Cannon. Using the triangle inequality to reduce the number of

comparisons required for similarity-based retrieval. In *SPIE vol. 2670 Storage and Retrieval for Still Image and*

Video Databases IV, pages 392–403, 1996.

[3] Tamara L. Berg, Alexander C. Berg, Jaety Edwards, Michael Maire, Ryan White, Yee whye Teh, Erik Leraned-

Miller, and David A. Forsyth. Names and faces in the news. In *Proc. Computer Vision and Pattern Recognition*

(CVPR04), pages II 848–854,Washington D.C., June 2004.

[4] Matthew Brand, Nuria Oliver, and Alex Pentland. Coupled hidden markov models for complex action recognition.

In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition CVPR97*, San Juan, Puerto Rico,

June 17–19 1997. IEEE Computer Society Press.

[5] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Region-based image querying. In *CVPR'97 Workshop on*

Content-Based Access of Image and Video Libraries, San Juan, Puerto Rico, June 1997.

[6] C. Carson and V. E. Ogle. Storage and retrieval of feature data for a very large online image collection. *Bulletin*

of the Technical Committee on Data Engineering, 19(4):19–27, December 1996.

[7] Robert T. Collins, Alan J. Lipton, and Takeo Kanade. A system for video surveillance and monitoring:vsam final

report. Technical Report CMU-RI-TR-00-12, CMU Robotics Institute, May 2000.

[8] Jonathan D. Courtney. Automatic video indexing via object motion analysis. *Pattern Recognition*, 30(4):607–

625, 1997.

[9] Ross Cutler and Larry S. Davis. Robust real-time periodic motion. *IEEE Trans. on Pattern Analysis and Machine*

Intelligence, 20(8):781–796, August 2000.

[10] James W. Davis and Aaron F. Bobick. The representation and recognition of action using temporal templates.

In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition CVPR97*, San Juan, Puerto Rico, June 17–19

1997. IEEE Computer Society Press.

[11] S. Edelman and D. Weinshall. A self-organizing multiple-view representation of 3d objects. *Biological Cybernetics*,

64:209–219, 1991.

[12] C. Faloutsos,W. Equitz,M Flickner,W. Niblack, D. Petkovic, and R. Barber. Efficient and effective querying by

image content. *Journal of Intelligent Information Systems*, 3:231–262, 1994.

[13] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised, scale-invarient learning. In

CVPR, pages 264–271, Minneapolis MN, July 2003.

[14] M. Flickner, H. Sawhney, and W. Niblack. Query by image and video content: The qbic system. *Computer*,

28(9):23–32, September 1995.

[15] D. A. Forsyth and M. M. Fleck. Finding people and animals by guided assembly. In *International Conference*

on Image Processing, 1997.

[16] D. A. Forsyth, J. Malik, M. M. Fleck, T. Leung, C. Bregler, C. Carson, and H. Greenspan. Finding pictures of

objects in large collections of images. In *Proc. of International Workshop on Object Recognition*, Cambridge,

MA, April 1996.

[17] Ismail Haritaoglu, David Harwood, and Larry S. Davis. W_ : Real-time surveillance of people and their activities.

IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(8):809–830, August 2000.

[18] S. Hongeng and R. Nevatia. Multi-agent event recognition. In *Proceedings of the 8th IEEE International Conference*

on Computer Vision, (ICCV2001), Vancouver, Canada, 2001.

[19] Chien-Yuan Huang and Octavia I. Camps. Object recognition using appearance-based parts and relations. In

IEEE Conference on Computer Vision and Pattern Recognition (CVPR97), pages 877–883, San Juan, Puerto

Rico, June 1997.

[20] Michal Irani and P. Anandan. Video indexing based on mosaic representation. *IEEE Trans. on Pattern Analysis*

and Machine Intelligence, 86(5):905–921, May 1998.

[21] Y. A. Ivanov and A. F. Bobick. Probabilistic parsing in action recognition. Technical Report 450, MIT Media

Lab, Vision and Modeling Group, 1997.

[22] C. E. Jacobs, A. Finkelstein, and D. H. Salesin. Fast multiresolution image querying. In *Proceedings of ACM SIGGRAPH-95*, pages 277–285, Los Angeles, CA, August 1995.

[23] Takeo Kanade, Robert T. Collins, Alan J. Lipton, Peter Burt, and Lambert Wixson. Advances in cooperative multi-sensor video surveillance. In *Proc. DARPA Image Understanding Workshop IUW98*, pages 3–24, Monterey, California, November 1998. DARPA.

[24] P. M. Kelly, M. Cannon, and D. R. Hush. Query by image example: The candid approach. In *SPIE Proc. Storage and Retrieval for Image and Video Databases III*, pages 238–249, 1995.

[25] Lily Lee, Raquel Romano, and Gideon Stein. Monitoring activities from multiple video streams: Establishing a common coordinate frame. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8), August 2000.

[26] Alan J. Lipton, Hironobu Fujiyoshi, and Raju S. Patil. Moving target classification and tracking from real-time video. In *IEEE Workshop on Applications of Computer Vision WACV*, pages 8–14, Princeton, NJ, October 1998.

[27] David Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, Korfu, Greece, 1999.

[28] T. P. Minka and R. W. Picard. Interactive learning using a society of models. *Pattern Recognition*, 30(4), April 1997.

[29] R. J. Morris and D. C. Hogg. Statistical models of object interaction. *International Journal of Computer Vision: IJCV*, 37(2):209–215, 2000.

[30] Hiroshi Murase and Shree K. Nayar. Learning and recognition of 3d objects from appearance. In *Proc. IEEE Workshop on Qualitative Vision*, pages 39–50, 1993.

[31] Randal Nelson and Isaac Green. Tracking objects using recognition. In *International Conference on Pattern Recognition (ICPR02), vol.2*, pages 1025–1039, Quebec City, Quebec, August 2002.

[32] Randal C. Nelson. Generating verbal descriptions of colored objects: Toward grounding language in perception. In *Proc. Workshop on the Application of Computer Vision (WACV 2005*, pages 46–53, Breckenridge Colorado, January 2005.

[33] Randal C. Nelson and Andrea Selinger. A cubist approach to object recognition. In *Proc. International Conference on Computer Vision (ICCV98)*, pages 614–621, Bombay, India, January 1998.

[34] Randal C. Nelson and Andrea Selinger. Large-scale tests of a keyed, appearance-based 3-d object recognition system. *Vision Research*, 38(15-16):2469–88, August 1998.

[35] V. Ogle and M. Stonebraker. Chabot: Retrieval from a relational database of images. *Computer*, 28(9):40–48, September 1995.

[36] Nuria M. Oliver, Barbara Rosario, and Alex P. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8), August 2000.

[37] R.W. Pentland, A. Picard and S. Sclaroff. Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3), June 1996.

[38] R.W. Picard and T. P. Minka. Vision texture for annotation. *ACM Journal of Multimedia Systems*, 3:3–14, 1995.

[39] Claudio Pinhanez and Aaron Bobick. Human action detection using pnf propagation of temporal constraints. In *Proc. of CVPR'98*, pages 898–904, Santa Barbara, California, June 1998.

[40] James M. Rehg, Maria Loughlin, and Keith Waters. Vision for a smart kiosk. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition CVPR97*, pages 690–696, San Juan, Puerto Rico, June 17–19 1997. IEEE Computer Society Press.

[41] C. Schmid and R. Mohr. Combining greyvalue invariants with local constraints for object recognition. In *Proc. CVPR96*, pages 872–877, San Francisco CA, June 1996.

[42] S. Sclaroff, L. Taycher, and M. LaCascia. Imagerover: A content-based image browser for the world wide web. In *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, June 1997.

[43] Andrea Selinger and Randal C. Nelson. Minimally supervised acquisition of 3d recognition models from cluttered images. In *Computer Vision and Pattern Recognition (CVPR01), Volume 1*, pages 213–220, Kauai, Hawaii, December 2001.

[44] J. R. Smith and S.-F. Chang. Visualseek: A fully automated content-based image query system. In *ACM Multimedia Conference*, Boston, MA, November 1996.

[45] Chris Stauffer and W. Eric L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on*

Pattern Analysis and Machine Intelligence, 22(8):747–757, August 2000.

[46] Matthew Turk and Alexander Pentland. Eigenfaces for recognition. *Journal of Neuroscience*, 3(1):71–86, 1991.

[47] Paul Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages 511–518, Kauai Hawaii, December 2001.

[48] Paul Viola andWilliam M.Wells. Alignment by maximization of information. *International Journal of Computer Vision*, 24(2):137–154, September 1997.

[49] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):780–785, August 1997.

[50] Christopher Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997.

[51] A. Ogale, A. Karapurkar and Y. Aloimonos, “View invariant recognition of action”, Proc. *Dynamic Vision Workshop*, ICCV 2005, Beijing, China.