International Journal of Wireless Communications, Networking and Mobile Computing
2014; 1(3): 20-28
Published online October 30, 2014 (http://www.aascit.org/journal/wcnmc)

AASCIT American Association for Science and Technology

Keywords
Object,
Image,
Pattern,
Action,
Activity,
Event,
Metadata

Received: October 08, 2014
Revised: October 16, 2014
Accepted: October 17, 2014

Internet based images and object activity recognition

Oladosu Olakunle Abimbola, Okikiola Folasade Mercy, Oladiboye Olasunkanmi Esther

Department of Computer Technology, Yaba College of Technology, Yaba, Lagos

Email address

kunledosu@gmail.com (Oladosu O. A.), sade.mercy@yahoo.com (Okikiola F. M.), estherworld2001@yahoo.com (Oladiboye O. E.)

Citation

Oladosu Olakunle Abimbola, Okikiola Folasade Mercy, Oladiboye Olasunkanmi Esther. Internet Based Images and Object Activity Recognition. *International Journal of Wireless Communications, Networking and Mobile Computing.* Vol. 1, No. 3, 2014, pp. 20-28.

Abstract

The availability of such imagery presents profound opportunities for image processing and computer vision research and exciting new applications in related fields. The challenge is how to organize, catalogue and retrieve such image data in a visually meaningful manner. It requires the development of new algorithms in the fields of scalable search in databases for global and/or local image features, integration with multi-modal data (meta-data), such as text and GPS information, information retrieval on the internet, scalable deployments of related algorithms on clustered computing architectures and multi-user interaction in social media. Object recognition has gained significant interest in scientific circles lately. Much early work on the analysis of activities took place within the computer vision community and leveraged video cameras as passive and non-invasive sensors. More recently, alternative paradigms based on dense sensors have emerged.

1. Introduction

Huge quantities of video and still imagery now exist on the web both in social media websites (Flicker, Facebook, and YouTube) and from webcams. Within these images and videos are stored the world's most significant sites, people, objects and events. Subjects are available from varying viewing position and angles, different times of day and night, changes in season, weather, and decade. The availability of such imagery presents profound opportunities for image processing and computer vision research and exciting new applications in related fields. The challenge is how to organize, catalogue and retrieve such image data in a visually meaningful manner. It requires the development of new algorithms in the fields of scalable search in databases for global and/or local image features, integration with multi-modal data (meta-data), such as text and GPS information, information retrieval on the internet, scalable deployments of related algorithms on clustered computing architectures and multi-user interaction in social media.

Object recognition has gained significant interest in scientific circles lately. This trend can be attributed mainly to two different reasons. First, spatio-temporal data derived from object motion is becoming more easily available due to advances in sensor technology and computing techniques. On the hardware side, advancements in sensor technology are resulting in low-cost versatile sensors. On the software side, advancements in computer vision have led to the design of robust object trackers that can handle occlusions, shape deformations and intensity changes in single- and multi-camera settings. Second, novel applications employing analysis of motion trajectory are

emerging due to enhanced interest in homeland security as well as due to prevalence of multimedia gadgets in commercial and scientific endeavors. Examples of the motion trajectory include tracking results from video trackers, sign language data measurements gathered from wired glove interfaces fitted with sensors, Global Positioning System (GPS) coordinates of satellite phones, cars using Car Navigation Systems (CNS), animal mobility experiments, etc. This spatio-temporal data embodies semantically rich information about the behavior of the object of interest, the action performed and the interaction among groups of objects .For example, in sign and gesture recognition, the signer moves his hands in specific pattern for a particular word. In sports video trajectory analysis and understanding can assist the players, coaches and sports analysts with strategies used on the field based on the motion patterns of players and their mutual interaction. Another important area is automatic video surveillance which is used, for example, in real-time observation of people and vehicles, in a busy environment, leading to a description of actions and mutual interactions. This application arises in scenarios as diverse as indoor and outdoor home and office scenes, railway and subway stations, parking lots, elevator and retail store videos, highway videos, etc. The complexity of the problem is exacerbated by low-resolution, weather-dependent video capture and the presence of multi-camera surveillance systems. The research challenge here is to quickly learn the permitted activities and set an alarm at any illegal or abnormal activity being performed. We emphasize that object motion plays the key role in the domain of activity analysis in general and in video surveillance in particular. Psychological studies have shown that human beings can routinely discriminate and recognize this kind of object motion using motion pattern, even in large viewing distances or poor visibility conditions; whereas, other cues such as clothes, appearance, or hair style tend to vanish at large distances or poor visibility conditions.

Nevertheless, developing high-accuracy activity classification and recognition algorithms using motion trajectories is still an extremely challenging task particularly when the number of activities to be recognized is relatively large. The object trajectory is typically modeled as a sequence of consecutive locations of the object on a coordinate system resulting in a vector in 2-D or 3-D Euclidean space. The measurement parameters, at each point in time, needed for object localization can be arbitrarily high-dimensional, distance (height or depth), silhouette of the object shape, and other metadata corresponding to object appearance and environment.

2. Review of Earlier Research Work

Much early work on the analysis of activities took place within the computer vision community and leveraged video cameras as passive and non-invasive sensors. More recently, alternative paradigms based on dense sensors have emerged. In this approach, tiny battery-free wireless sensors are attached to objects and surfaces in a space and can provide direct measurements of the user's proximity to objects and regions of the environment. For example, described a system based on (radio frequency identification) RFID tags for recognizing a subset of the activities of daily living, a canonical activity recognition task for computer-assisted care applications. While dense sensors are appealing due to their low cost and simplicity, they have several drawbacks in comparison to video-based analysis. First they require objects and often people to be instrumented. Second, they do not work with certain types of objects such as metallic objects, food items, and objects that are very small. In addition there are problems with signal drop out, latency, and confusion between labels during reading.

Many activity recognition methods in computer vision have focused on the representation and modeling of actions, which are the atomic units within activities. This line of work explores tracking methods and other forms of spatio-temporal video analysis in order to sense what the actor is doing. In other words, they attempt to identify the activity by sensing the verbs. A common theme in these works is the exploration of spatio-temporal video features. Other work has addressed the use of multiple resolutions. Temporal constraints on actions are addressed either through the use of probabilistic models such as HMM's or SCFG's, or through explicit temporal correlation methods. The recognition of actions can often be aided by incorporating context from the environment. We loosely characterize these approaches as sensing verbs plus context.

In the work of (Peursum, West, & Venkatesh, 2005) actions can be discriminated by identifying the spatial location within the scene in which they occur. Hand motions which might be ambiguous in general can be correctly classified as typing when they occur in the vicinity of the keyboard. The W4 system used outdoor scene context in conjunction with a robust blob-tracking algorithm to analyze scenarios in which multiple people interacted and exchanged bags and other objects. Other sources of task-specific context include the identification of roads and entrances/exits in parking lot surveillance and tracking the components of a blood glucose monitor. In contrast to these works, we are interested in domains where the action and spatial location are of limited utility in recognizing activities.

(Bao & Intille, 2004) noted that most previous studies examining activity recognition from accelerometer data were not suitable for real-world situations and were conducted either in laboratory conditions or used limited datasets. They assessed the performance of algorithms in identifying twenty activities under semi-naturalistic, simulated real-world conditions using five biaxial accelerometers. Decision table, distance-based learning, decision tree and Naive Bayes classifiers were used with providing the best performance recognizing everyday activities with an overall accuracy of 84%. The above study also identified the optimal single accelerometer position, for the set of activities they chose, as

being on the thigh and that accuracy increased by 25% by using more than one accelerometer. It was shown that acceleration data could be augmented with heart rate data to determine the intensity of physical activities. Pirttikangas et al undertook a study using coin-sized sensor devices attached to four parts of the body: right thigh and wrist, left wrist and a necklace. 17 daily activities were examined using triaxial accelerometer and heart rate data. Two classifiers were used (multilayer perceptrons and kNN classifiers) with kNN achieving a 90.61% aggregate recognition rate for 4-fold cross-validation. Interestingly, heart rate data was collected but not used in the activity recognition process.

A major challenge in this approach is the need for a robust general-purpose object recognition system which could reliably discriminate between hundreds of different cooking items under real-world imaging conditions. Building models for object recognition usually requires labeled training images without a cluttered background. In order to obtain this, a significant amount of work (segmentation and labeling of objects) is required. In contrast, our work leverages temporal continuity in video frames to roughly segment the moving object. An object is modeled as a bag of SIFT features and learning object models is equivalent to assigning the probabilities of seeing different SIFT features in an object. Our approach is similar to which assigned features from independent images into ‘topics’ using LSA, an unsupervised learning method. Their results showed that the revealed topics usually coincided with objects. In contrast, we use sparse and noisy RFID measurements to guide the learning process.

Recently, dense sensors have been proposed as an alternative to vision-based object recognition for obtaining object information in activity recognition tasks. In these works, wireless sensors attached to both humans and objects make it possible to directly measure actor-object interactions. Possible sensor data includes the identities of people and objects (e.g. RFID sensors) as well as their position sand velocities (e.g. accelerometers and audio sensors).

Activity recognition fits into the bigger domain of context awareness by making devices aware of the activity or activities of the user. The ability to recognize human activities is a key factor if computing systems are to interact seamlessly with the user’s environment. Context awareness is leading to the ’reinvention’ of some domains such as healthcare with studies examining a diverse range of applications such as hospital worker activity estimation , chronic disease management and remote patient monitoring.

In context aware computing, data can be collected from a diverse range of sensors such as audio sensors, image sensors and accelerometers. Accelerometers facilitate the real-time recording of acceleration data along the x-, y- or z-axis. Due to their ever-diminishing size and embeddable nature, accelerometers can be unobtrusively worn by users. It has been noted that accelerometers have successfully crossed over to the mainstream via devices such as Apple’s iPhone.

Much recent research has applied classification algorithms to accelerometer data in order to increase activity recognition accuracy with some commentators stating that activity recognition is primarily a classification problem. Two classifiers, k-NN and J48/C4.5 (J48 is the Weka Toolkit Java implementation of C4.5), were evaluated in this study. The Weka Toolkit is a collection of state-of-earth machine learning algorithms and data pre-processing tools developed at the University of Waikato in New Zealand. Lombriser et al identify k-NN and J48/C4.5 as being “the classifiers with the least complexities but rendering acceptable performance.

The recognition of actions can often be aided by incorporating context from the environment. We loosely characterize these approaches as sensing verbs plus context. For example, in the work of Moore et al. and Peursum et al, actions can be discriminated by identifying the spatial location within the scene in which they occur. Hand motions which might be ambiguous in general can be correctly classified as typing when they occur in the vicinity of the keyboard. The W4 system used outdoor scene context in conjunction with a robust blob-tracking algorithm to analyze scenarios in which multiple people interacted and exchanged bags and other objects.

Hierarchical architectures have been shown to outperform single-template (flat) object recognition systems on a variety of object recognition tasks (e.g., face detection and car detection. In particular, constellation models have been shown to be able to learn to recognize many objects (one at a time) using an unsegmented training set from just a few examples. Multilayered convolution networks were shown to perform extremely well in the domain of digit recognition and, more recently, generic object recognition and face identification. The simplest and one of the most popular appearance based feature descriptors corresponds to a small gray value patch of an image, also called component, part or fragment. Such patch-based descriptors are, however, limited in their ability to capture variations in the object appearance: They are very selective for target shape but lack invariance with respect to object transformations. At the other extreme, histogram-based descriptors have been shown to be very robust with respect to object transformations. Perhaps the most popular features are the SIFT features, which excel in the redetection of a previously seen object under new image transformations and have been shown to outperform other descriptors.

However, as we confirmed experimentally, with such a degree of invariance, it is very unlikely that these features could perform well on a generic object recognition task. The new appearance-based feature descriptors described here exhibit a balanced trade-off between invariance and selectivity. They are more flexible than image patches and more selective than local histogram-based descriptors. Though they are not strictly invariant to rotation, invariance to rotation could, in principle, be introduced via the training set (e.g., by introducing rotated versions of the original input). Much early work on the analysis of activities took place within the computer vision community and leveraged video cameras as passive and non-invasive sensors. More recently, alternative paradigms based on dense sensors have emerged.

In this approach, tiny Battery-free wireless sensors are attached to objects and surfaces in a space and can provide direct measurements of the user's proximity to objects and regions of the environment. For example, described a system based on RFID tags for recognizing a subset of the activities of daily living, a canonical activity recognition task for computer-assisted care applications. While dense sensors are appealing due to their low cost and simplicity, they have several drawbacks in comparison to video-based analysis. First they require objects and often people to be instrumented. Second, they do not work with certain types of objects such as metallic objects, food items, and objects that are very small.

In addition there are problems with signal drop out, latency, and confusion between labels during reading, many activity recognition methods in computer vision have focused on the representation and modeling of actions, which are the atomic units within activities. This line of work explores tracking methods and other forms of spatio-temporal video analysis in order to sense what the actor is doing. In other words, they attempt to identify the activity by sensing the verbs. A common theme in these works is the exploration of spatio-temporal video features,

Other work has addressed the use of multiple resolutions. Temporal constraints on actions are addressed either through the use of probabilistic model such as HMM's or SCFG's, or through explicit temporal correlation methods. The recognition of actions can often be aided by incorporating context from the environment. We loosely characterize these approaches as sensing verbs plus context. In contrast to these works, we are interested in domains where the action and spatial location are of limited utility in recognizing activities. Many different cooking activities, for example, involve picking up and putting down objects within a single counter-top area. To differentiate among these activities it is necessary to identify the objects which are being manipulated. We characterize this approach as recognizing activities by sensing the object use. The cooking domain involves a large number of different objects which are shared across multiple activities and are not restricted to any particular location in the image. A major challenge in this approach is the need for a robust general-purpose object recognition system which could reliably discriminate between hundreds of different cooking items under real-world imaging conditions. Building models for object recognition usually requires labeled training images without a cluttered background. In order to obtain this, a significant amount of work (segmentation and labeling of objects) is required. In contrast, our work leverages temporal continuity in video frames to roughly segment the moving object. An object is modeled as a bag of SIFT features and learning object models is equivalent to assigning the probabilities of seeing different SIFT features in an object. Their results showed that the revealed topics usually coincided with objects. In contrast, we use sparse and noisy RFID measurements to guide the learning process.

Recently, dense sensors have been proposed as an alternative to vision-based object recognition for obtaining object information in activity recognition tasks. In these works, wireless sensors attached to both humans and objects make it possible to directly measure actor-object interactions. Possible sensor data includes the identities of people and objects (e.g. RFID sensors) as well as their positions and velocities (e.g. accelerometers and audio sensors). In RFID-based systems, activities are represented as probability distributions over sequences of object-use obtained from sparse and noisy RFID readings. In other approaches, information from accelerometers is used to identify actions such as walking and climbing stairs. RFID and vision were also used as complementary sensors. In RFID and vision were used to track object and human independently, and were combined using rules. In contrast to this latter work, utilization of RFID sensors to fit vision object models in an integrated DBN framework.

Another aspect of activity recognition which has received significant attention is computational models of activity which can serve as a constraint on the interpretation of noisy sensor data. Complex activities such as baking a cake can be decomposed into subtasks, and constraints from the domain (e.g. the oven must be preheated before it can be used) result in partial orderings of these subtasks. There has been much interesting work in representing and exploiting these constraints during recognition. Although RFID can sense the use of objects, in practice it has several limitations which motivate us to bootstrap the RFID readings using vision. If the bracelet is close to an object by accident, it may indicate erroneously that the object is being manipulated. If a tagged object is grasped far from the tag, on the other hand, the manipulation may be missed.

This section provides a survey of the related work from recent literature in the areas of trajectory representation, statistical modeling and applications of trajectory-based representation and learning. Studies into human psychology have shown the extra-ordinary ability of human beings to recognize object motion even from minimal information system such as Moving Light Displays (MLDs). Such displays are obtained by making a video of moving subjects wearing reflective pads/light bulbs on their body joints in almost dark conditions. In spite of the paucity of information, human observers easily perceive not only motion but also the kind of motion; e.g., walking, running, dancing, cycling, etc. Based on this understanding, object motion has been an important feature for the representation and discrimination of one object from another in video applications. Earlier approaches in motion-based methods focused on object tracking from raw and compressed domain videos. Indexing and searching based on object motion as the dominant cue has attracted a lot of research activity in the past few years.

Chen et al. segment each trajectory into sub trajectories using fine-scale wavelet coefficients at high levels of decomposition. A feature vector is then extracted from each sub trajectory comprising of features like acceleration, velocity, sub trajectory length, etc. Distances between each sub trajectory in query trajectory and all the indexed sub trajectories are computed to generate a list of similar

trajectories in the database. This approach suffers from the fact that the representation is based on adhoc features which are not tolerant to affine transformations of the trajectories. Also, the feature vectors lie in a non-uniform space, so the matching process has to compute the overall distance based on weighted average of individual features. Previous work on trajectory indexing and retrieval segments the trajectories based on dominant sign changes in curvature data. We represent the sub trajectories using PCA coefficients.

The view-invariant representation of trajectories for scenarios where similar trajectories are captured from different viewpoints. View-invariant representation has also been addressed in for modeling and recognizing actions performed by individuals in video sequences. The representation is based on dynamic instants (segmentation points) of the trajectories. For each dynamic instant in the trajectory, frame number, location of the hand and 'sign' of the instant (-ve for counter clockwise turn and +ve for clockwise turn) is stored. The matching is performed on trajectories with the same number of dynamic instants and same sign permutations. This approach, though compact in representation, cannot be used for partial trajectory processing or generic trajectory representation.

Yacoob et.al. have presented a framework for modeling and recognition of human motions based on principal components. Each activity is represented by eight motion parameters recovered from five body parts of the human walking scenario. In a semantic event detection technique for snooker videos is presented. Trajectory of the while ball is generated using a color-based particle filter. The implementation of the particle filter allows for ball collision detection and ball pot detection. A separate ball track is instantiated upon detection of a collision and the state of the new ball can be monitored. The evolution of the white ball position is modeled using a discrete HMM. In the issue of recognizing a set of plays from American football videos is considered. Using a set of classes each representing a particular game plan and computation of perceptual features from trajectories, the propagation of uncertainty paradigm is implemented using automatically generated Bayesian network. The problem with above approaches is that they are highly domain-dependant, with domain knowledge and sensor dependence on video data being intimately woven into the systems. A sensor-independent approach towards modeling activity performed by a group of objects (persons, cars, etc.) is presented. Objects in scene are taken as points and they consider the 'shape' formed by a configuration of point objects at a given time instant. This 'shape' is tracked over time, normal shape is learnt and abnormality is detected as perturbation in this shape. Although robust for multi-agent abnormal activity detection, this approach cannot be applied for single object trajectories.

Vinciarelli et al have used PCA and ICA (Independent Component Analysis) along with HMMs for word recognition in hand writing recognition application. De la Torre et al use PCA and HMM for tracking and recognition for lip-tracking and eye-tracking. Martin et al model the trajectories for gesture recognition using multidimensional histogram of gestures. In their approach, no segmentation to obtain sub trajectories is performed; only the recent history is taken into account. The results are reported in terms of head movements for two gestures of 'Yes' and 'No'; four single stroke letters from graffiti characters five expressions for facial expression analysis from gray scale images of size 44x60. Starner and Pentland address the issue of American Sign Language recognition from video sequences. An 8-element feature vector is obtained consisting of each hand's x and y positions, angle of axis of least inertia, and eccentricity of bounding ellipse is used.

Bettinger et al address the problems of learning a person's facial behaviors from video sequences and synthesizing sequences demonstrating the same behavior. A sequence of a face is represented as a parameter sequence labeled as a trajectory in parameter space, which is then segmented into sub trajectories. HMMs are then trained on this data to learn the facial behavior models. It is important to point out that the notion of trajectory, the process of segmentation and representation used in are entirely different than the method presented.

3. Methodology

Firstly, we attempt to classify events in static images by integrating scene and object categorizations. Our goal is to classify the event in the image as well as to provide a number of semantic labels to the objects and scene environment within the image. For example, given a rowing scene, our algorithm recognizes the event as rowing by classifying the environment as a lake and recognizing those critical objects in the image as athletes, rowing boat, water, etc. We achieve this integrative and holistic recognition through a generative graphical model.

4. Image and Object Recognition Process

This is the task of finding a given object in an image or video sequence. Humans recognize a multitude of objects in images with little effort, despite the fact that the image of the objects may vary in different viewpoints, in many different sizes / scale or even when they are translated or rotated. Objects can even be recognized when they are partially obstructed from view.

Object recognition is the process whereby observers are able to recognize three-dimensional objects despite receiving only two-dimensional input that varies greatly depending on viewing conditions.

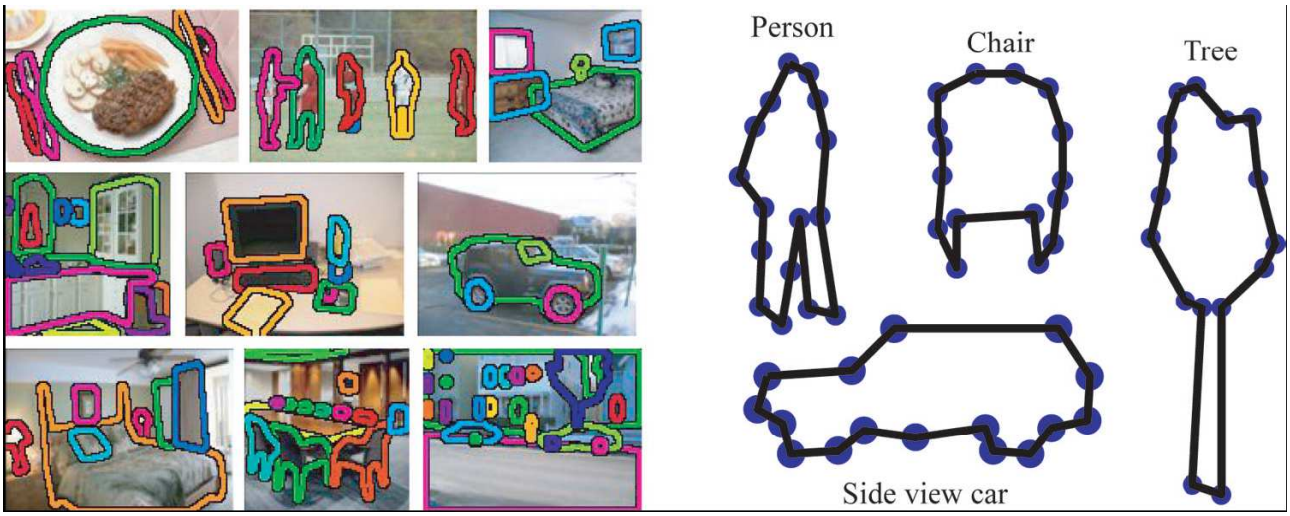


Figure 1. *Image showing object recognition*

Object recognition can be described as all of the following:

- Artificial intelligence
- Application of artificial intelligence
- Pattern recognition

Appearance-based object recognition methods have recently demonstrated good performance on a variety of problems. However, many of these methods either require good whole-object segmentation, which severely limits their performance in the presence of clutter, occlusion, or background changes; or utilize simple conjunctions of low-level features, which cause crosstalk problems as the number of objects is increased. We are investigating an appearance-based object recognition system using a keyed, multi-level context representation that ameliorates many of these problems, and can be used with complex, curved shapes. Pictures on this page are from a training database we have used in system tests. The basic idea is to represent the visual appearance of an object as a loosely structured combination of a number of local context regions keyed by distinctive key features, or fragments. A local context region can be thought of as an image patch surrounding the key feature and containing a representation of other features that intersect the patch. Now under different conditions (e.g. lighting, background, changes in orientation etc.) the feature extraction process will find some of these distinctive keys, but in general not all of them. Also, even with local contextual verification, such keys may well be consistent with a number of global hypotheses. However, the fraction that can be found by existing feature extraction processes is frequently sufficient to identify objects in the scene, once the global evidence is assembled. This addresses one of the principle problems of object recognition, which is that, in any but rather artificial conditions, it has so far proved impossible to reliably segment whole objects on a bottom-up basis. In the current system, local features based on automatically extracted boundary fragments are used to represent multiple 2-D views (aspects) of rigid 3-D objects, but the basic idea could be applied to other features and other representations.

4.1. Applications of Object Recognition

Object recognition methods have the following applications:

- Android Eyes - Object Recognition
- Image panoramas
- Image watermarking
- Face detection
- Optical Character Recognition
- Manufacturing Quality Control
- Content-Based Image Indexing
- Object Counting and Monitoring
- Automated vehicle parking systems
- Video Stabilization

4.2. Object Classification

A classifier is an algorithm that takes a set of parameters (or features) that characterize objects (or instances) and uses them to determine the type (or class) of each object. The classic example in astronomy is distinguishing stars from galaxies. For each object, one measures a number of properties (brightness, size, etc.); the classifier then uses these properties to determine whether each object is a star or a galaxy. Classifiers need not give simple yes/no answers -- they can also give an estimate of the probability that an object belongs to each of the candidate classes.

Techniques thus far only classify objects based on their shape, color, texture, etc. These are only representative of the light reflected by an object. Humans classify objects in many ways, including an object's function.

4.3. Classification Method

Object classification methods are very useful tools for data exploration in large, complex problems. Such tools have traditionally been described as artificial intelligence methods, which may account for some of the skepticism among astronomers as to the applicability of these methods to quantitative analysis. Classifiers need not be seen as

mysterious black boxes that just spit out the answers; decision trees, in particular, represent a relatively simple geometric partitioning of the parameter space that can provide an accurate, understandable characterization of a complex data set. The classification problem becomes very hard when there are many parameters. There are so many different combinations of parameters that techniques based on exhaustive searches of the parameter space are computationally infeasible. Practical methods for classification always involve a heuristic approach intended to find a ``good-enough'' solution to the optimization problem. The classification methods are:

a) Neural Networks
b) Nearest-Neighbor Classifiers
c) Decision

4.4. Activity Recognition

Activity recognition aims to recognize the actions and goals of one or more agents from a series of observations on the agents' actions and the environmental conditions. An activity recognition approach based on object use can be particularly useful in domains such as cooking, which involve a relatively small number of repeated actions such as chopping, pouring, spreading, etc. Object use information can help discriminate between activities such as making toast and making a sandwich, which may be similar from the standpoint of the actions alone. Such distinctions can be important for application domains such as health monitoring or memory aids. A significant issue in the development of an object-based approach is its scalability, given the potentially large number of objects that must be discriminated, and the difficulty of obtaining labeled training data for each object under realistic conditions. Potential users are unlikely to be willing to spend a significant amount of time training a recognition system by presenting it with individually-labeled object instances. However, given video of everyday household activities, it is possible that object models could be extracted automatically if a sufficiently informative training signal was available.

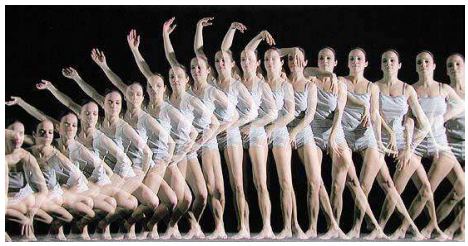
A method for activity recognition based upon automatically-acquired models of activities and the objects that they involve. In other words, we recognize activities by identifying the objects which are being used in the image. An activity recognition approach based on object use can be particularly useful in domains such as cooking, which involve a relatively small number of repeated actions such as chopping, pouring, spreading, etc.

Actions can be represented in the following ways.

- Categories: Walking, hammering, dancing, skiing, sitting down, standing up, and jumping.
- Poses
- Nouns and Predicates: <man, swings, hammer> <man, hits, nail, w/ hammer>

Actions can also be identify in

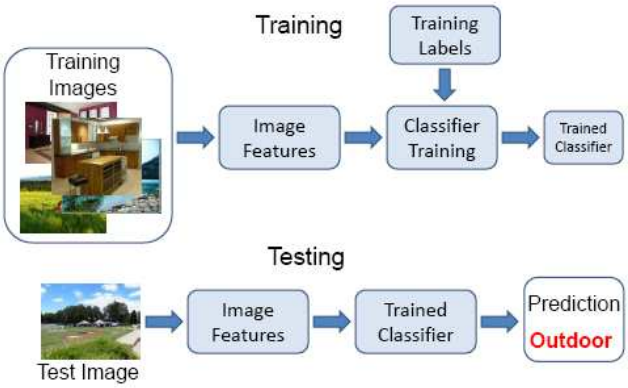
1 Motion

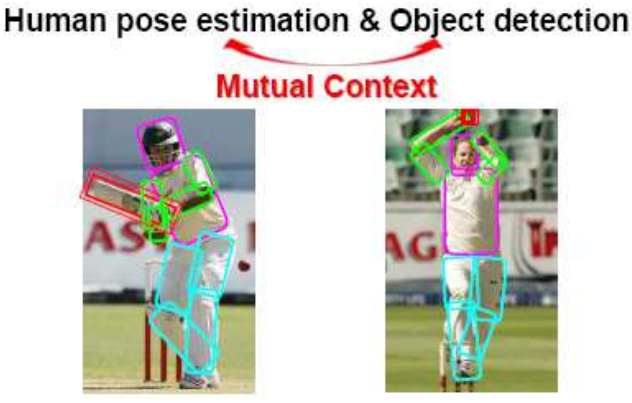


2 Pose

3 Held object

4 Image Categorization





4.5. Types of Activity Recognition

Sensor-based, single-user activity recognition: - integrates the emerging area of sensor networks with novel data mining and machine learning techniques to model a wide range of human activities. Mobile devices (e.g. smart phones) provide sufficient sensor data and calculation power to enable physical activity recognition to provide an estimation of the energy consumption during everyday life. Sensor-based activity recognition researchers believe that by empowering ubiquitous computers and sensors to monitor the behavior of agents (under consent), these computers will be better suited to act on our behalf.

a) Sensor-based, multi-user activity recognition: - Recognizing activities for multiple users. Other sensor technology such as acceleration sensors were used for identifying group activity patterns during office scenarios.
b) Vision-based activity recognition: - It is a very important and challenging problem to track and understand the behavior of agents through videos taken by various cameras. The primary technique employed is computer vision. Vision-based activity recognition has found many applications such as human-computer interaction, user interface design, robot learning, and surveillance, among others. Scientific conferences where vision based activity recognition work often appear are ICCV and CVPR.

5. Pattern Matching

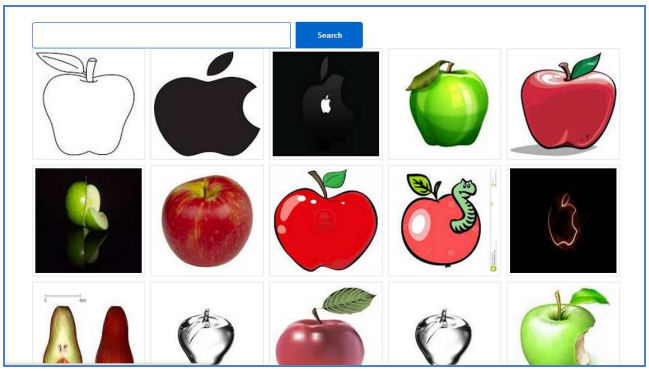


Figure 2. *A. Pattern Matching*

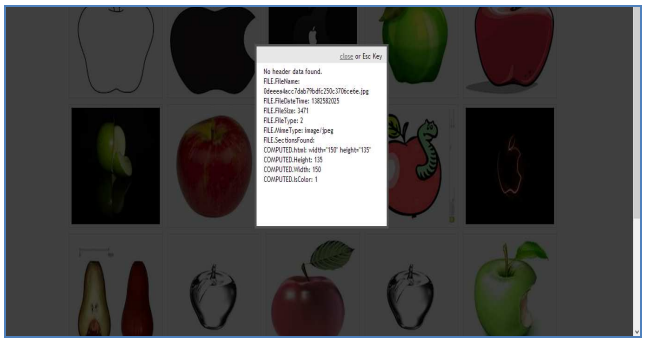


Figure 2. *B. Pattern Matching*

6. Summary and Conclusion

To recognize an object, that is to answer the question "what object is in this image?" key features together with their local contexts are extracted from the image, and fed into the associative memory. All matches are retrieved, and for each match, the associated information is used to compute a hypothesis about the identity, view, and configuration of a possible object. If any matches are found, the evidence associated with them is updated to reflect the new information.

Techniques thus far only classify objects based on their shape, color, texture, etc. humans classify objects many ways, including an object's function. Computer object recognition techniques lack some abilities which are simple for humans. Some work done, but it is just the beginning of exploring the problem. So far actions are mainly categorical and most approaches are classification using simple features.

Modern object recognition techniques can provide much functionality in controlled environments. Simulation of human object recognition capabilities is a long way off. The basic recognition strategy is to utilize a database (here viewed as an associative memory) of key features embedded in local contexts, which is organized so that access via an unknown key feature evokes associated hypotheses for the identity and configuration of all known objects that could have produced such an embedded feature.

A fundamental component of the approach is the use of distinctive local features we call *keys*. A key is any robustly extractable part or feature that has sufficient information content to specify a configuration of an associated object plus enough additional, pose-insensitive (sometimes called semi-invariant) parameters to provide efficient indexing. The local context amplifies the power of the feature by providing a means of verification. To find an object of known characteristics in a scene, that is to answer the question of the form "where is the dog in this image?", the same procedure is followed, except that key feature matches are filtered on the basis of whether the came from a view of a dog. This actually provides a rather powerful mechanism for partially indexed retrieval, since the filtering can occur on any combination of attributes that we care to associate with the features, either in the database, or from the image.

References

[1] Andrea Selinger and Randal C. Nelson, ``A Perceptual Grouping Hierarchy for Appearance-Based 3D Object Recognition", Computer Vision and Image Understanding, vol. 76, no. 1, October 1999, pp.83-92

[2] Breiman, L. 1996, Machine Learning, 24(2), 123

[3] B. Schiele and J. L. Crowley "Recognition without correspondence using multidimensional receptive field histograms", International Journal of Computer Vision, 36:1, 31-50, 2000

[4] Bülthoff, H. & Edelman, S. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Natl. Acad. Sci. USA* 89, 60–64 (1992).

[5] Derek Hao Hu, Qiang Yang. "CIGAR: Concurrent and Interleaving Goal and Activity Recognition", to appear in AAAI 2008

[6] Favela, Monica Tentori, Luis A. Castro, Victor M. Gonzalez, Elisa B. Moran, and Ana I.

[7] Mart´ınez-Garc´ıa. Activity recognition for context-aware hospital applications: issues and opportunities for the deployment of pervasive networks. Mob. Netw. Appl., 12(2-3):155–171, 2007.

[8] http://csdl.computer.org/comp/proceedings/icip/1997/8183/03/81830408abs.htm

[9] Jie Yin, Dou Shen, Qiang Yang and Ze-nian Li "Activity Lecture Notes in Computer Science, pages 1–17, 2004.

[10] L. Bao and S. Intille. Activity recognition from user annotated acceleration data.

[11] Ling Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. In

[12] Pervasive 2004, pages 1–17. Springer, 2004.

[13] Mel, B. SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Compuet.* 9, 777–804 (1997).

[14] Michael Tarr, Brown University http://www.cog.brown.edu/~tarr/pdf/Tarr02ECS.pdf#search='object%20recognition

[15] M. J. Swain and D. H. Ballard "Colour indexing", International Journal of Computer Vision, 7:1, 11-32, 1991.

[16] P. Peursum, G. West, and S. Venkatesh. Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In Proc. ICCV, volume 1, pages 82–89, 2005.

[17] Recognition through Goal-Based Segmentation". Proceedings of the Twentieth National

[18] Recognition. In Proc. of the 7th Annual IEEE International Conference on Pervasive Computing and Communications (Percom '09), Galveston, Texas, March 9–13, 2009.

[19] Randal C. Nelson, ``Memory-Based Recognition for 3-D Objects", Proc. ARPA Image Understanding Workshop, Palm Springs CA, February 1996, 1305-1310.

[20] Wallis, G. & Rolls, E. A model of invariant object recognition in the visual system Programming. Neurobiol. 51, 167–194 (1997).