A Framework for Designing Event Detectors

A Framework for Designing Event Detectors

Niels Haering and Niels da Vitoria Lobo

Abstract

We propose an intermediate represenation of video data to simplify the design of event detectors. This approach allows the description of complex spatio-temporal actions on objects in terms of measurable image and video properties, such as color, texture, spatio-temporal patterns, motion, and shot boundary information. We show the usefulness of the color and texture measures for the recognition of natural objects, and the combination of color, texture, motion, and shot boundary information for the detection of hunt events in wildlife documentaries. The proposed approach allows the annotation of video data with both low level color, texture, and motion pattern tags, as well as higher level object and event information. A description of video data in terms of the derived primitives reduces the gap between semantic and syntactic descriptions of events, thus simplifying the design of event detection methods. We conclude this report by suggesting solutions for a number of sample events in terms of the proposed intermediate representation.

University of Central Florida Technical Report CS-TR-99-01

January, 1999

For additional copies, write:

Technical Reports
School of Computer Science
University of Central Florida
Orlando, FL. 32816
U.S.A.

I. Introduction

We propose a rich intermediate representation of video data to simplify the design of event detectors.

Consider the following examples of events: wildlife hunts, kissing, explosions, overtaking on the highway, turning a corner, picking an orange, feeding the cat, buying a car, surfing, entering a building, rock-climbing, a car crash, etc.

For many of these events we do not need complete object or motion descriptions in order to establish their occurrence. For wildlife hunts it would be difficult to model the terrain, the predators' shape, motivation, and health, the occlusions, the lighting conditions, the camera parameters, the effects of the video compression scheme on the objects, etc. We might not care about the kind of predator, the predator's limp due to a thorn in its foot, the prey, the prey's speed or global motion parameters, the weather, the camera, the compression scheme, etc.

Likewise, we may not know or need to know the people who are kissing; we may not know or need to know what exploded; we may not care about the absolute or relative speed and the kind of vehicles that are involved in the overtaking manouver, etc.

However, there are spatial, temporal, and spatio-temporal patterns that are significant, and without which the detection of an event would be very difficult. While we may have a good idea about what it is that defines an event in terms of objects and actions it may be difficult to specify how to detect the objects and actions. Many actions can operate on a number of objects. For instance, to recognize a running action we need not know whether the running object is a human, an animal, or a cartoon character. Similarly, to recognize a zebra we do not need to know if it is running, sleeping, or feeding.

We propse a rich intermediate representation of video data to narrow the gap between abstract event descriptions and pixel intensities. For this purpose we derive spatial features from color and texture measures, motion measures from the locations of corresponding image regions in multiple frames, qualitative motion information from spatio-temporal measures. Since shots are natural building blocks of events we also gather shot boundary information and maintain statistics for regions of interest throughout shots. Spatial, temporal, spatio-temporal frame measures together with a range of shot characteristics provide a rich representation of video content that simplifies the design of event detectors.

A. Rich Image Descriptions

Rich descriptions of the world are known to simplify classification and recognition tasks, since the potential of simple decision surfaces that separate the classes increases monotonously with the dimensionality of the input representation. If a problem is, linearly separable in a k dimensional representation it is guaranteed to be linearly separable if further dimensions are being used to describe the problem. Perhaps more interesting is the fact that a problem that is only non-linearly separable in a k dimensional representation may be linearly separable if further dimensions are beeing added to describe the problem. This is important, since simpler decision surfaces often produce better classifications, as noted by Sir Occam around 1325 A.D.: "The simplest explanation is the best".

Note also that this fact is not dependent on orthogonality between existing and additional dimensions. The greater discernibility of patterns in higher dimensional problem representations facilitates the description of more complex semantic spatio-temporal events of interest.

B. Constructing Visual Primitives

Most work in object recognition and event detection aims to solve fairly complex tasks in simple environments. We suggest an alternative approach that

1. constructs an internal representation of visual data,
2. creates simple object and event detectors based on this internal representation, and
3. constructs more complex events in terms of both the internal representation, as well as the simple object and event detectors.

C. Object Recognition

An insightful definition of the goal of object recognition was given at the Workshop on 3D Object Representation for Computer Vision [32] in 1994: "*... model-based vision must go beyond pose estimation and into actual object recognition: for model databases containing thousands of objects, we cannot afford to try every model, estimate its pose, then verify its presence in the image using the estimated pose. We must also tackle the difficult problems of extracting the relevant information from images (segmentation), automatically constructing the object models, indexing in sub-linear time the model database and eventually integrating the corresponding modules into working end-to-end recognition systems.*"

The goal of Computer Vision research is to achieve complex tasks in unrestricted environments.

But since this was too hard researchers started off, with simple tasks in heavily constrained environments. If a reasonable complex task had been achieved the constraints on the environment were relaxed. This violated the initial assumptions, and the approaches were rendered ineffective. Usually large chunks of the programs had to be reprogrammed. The methods developed for the block world example, mentioned above, can not be used for natural objects or for human made objects that do not have well defined corners; but even the assumption that corner locations can be obtained robustly and automatically proved too difficult.

An analogous problem in robotics research prompted Rodney Brooks [5] to propose a subsumption architecture for the construction of robots. Rather than starting off with complex tasks in simple environments, he suggested to start off with simple tasks in environments of full complexity.

This approach significantly motivated the work described in this report. We propose an end-to-end object recognition approach that uses rich image descriptions as intermediate representations. The image descriptors consist of color and texture measures describing different properties of image regions. While each descriptor in isolation is weak, the combination of a number of them achieves robust object recognition. This approach performs well under a wide range of lighting and imaging conditions, object sizes and orientations, image compression and transformation schemes as well as significant shape variations in the objects. Many non-rigid objects, like sky/clouds, trees, grass, fire, water, rocks, mountains, etc. cannot make use of geometry or shape based recognition schemes. We found this approach to be effective for the classification of natural scenes into the categories trees, grass, sky/clouds, rock, and animal.

A synthetic approach to object and event recognition reduces the burden on the verification process, by limiting the number of possible models (interpretations) for a given image or video segment. It is an important pre-requisite for a high level knowledge based search that selects the most relevant model given both bottom-up evidence and top-down confirmation.

D. Event Recognition

In many ways event recognition is to video data what object recognition is to image data. In order to describe the content of image data, it is necessary to be able to detect and recognize objects, while for video data it is necessary that we can detect and recognize actions and events, such as chasing, e.g. [**?**], entering a room, e.g. [10], explosions, e.g. [22], etc.. Although the human visual system can often infer events, like those mentioned above, even from still images, it

is generally believed that event detection or recognition from video data is simpler. For example, from a single image of a human made satellite in space we cannot conclude whether it is plunging to earth, heading for Mars, or whether it is orbiting earth, while a video sequence of the satellite can eliminate this uncertainty.

We will limit ourselves to describing events that consist of actions on objects over time (as opposed to recognizing events from still images).

E. Previous Work

The amount of image and video information that can be accessed and consumed from people's living rooms has been ever increasing. This trend may be further accelerated due to the convergence of both technology and functionalities supported by future television receivers and personal computers. To obtain the information of interest tools are needed to help users to extract relevant content and to effectively navigate through the large amount of available image and video information.

Existing content-based image and video indexing and retrieval methods may be classified into the following three categories: (1) syntactic structurization; (2) image or video categorization; and (3) extraction of semantics.

For *image retrieval*, work in the first category has concentrated on color, texture, measures of the entire images and shape descriptions of selected objects [36], [42]. Work in the second category has focused on categorizing pictorial data into graphics, logos or images [3]. Work in the third category aimed at classifying images based on their semantic content, e.g. indoor-outdoor [41], landscape-cityscape [43].

For *video retrieval*, work in the first category has concentrated on (a) shot boundary detection and key frame extraction, e.g., [2], [49]; (b) shot clustering, e.g., [47]; (c) table of content creation, e.g., [14]; (d) video summarization, e.g., [28]; and (e) video skimming [39]. These methods are in general computationally simple and their performance is relatively robust. Their results, however, may not necessarily be semantically meaningful or relevant since they do not attempt to model and estimate the semantic content of the video. For consumer oriented applications, semantically irrelevant results may distract the user and lead to frustrating search or browsing experience. The work in the second category tries to classify video sequences into categories such as news, sports, action movies, close-ups, crowd, etc. [24], [44]. These methods provide classification results which may facilitate users to browse video sequences at a coarse level. Video content analysis

at a finer level is probably needed, to more effectively help users find what they are looking for. In fact, consumers often express their search items in terms of more exact semantic labels, such as keywords describing objects, actions, and events. The work in the third category has been mostly specific to particular domains. For example, methods have been proposed to detect events in (a) football games [23]; (b) soccer games [48]; (c) basketball games [38]; (d) baseball games [26]; and (e) sites under surveillance [10]. The advantages of these methods include that the detected events are semantically meaningful and usually significant to users. The major disadvantage, however, is that many of these methods are heavily dependent on specific artifacts such as editing patterns in the broadcast programs, which makes them difficult to extend for the detection of other events. A query-by-sketch method has also been proposed recently in [7] to detect motion events. The advantage of this method is that it is domain-independent and therefore may be useful for different applications. For consumer applications, however, sketching needs cumbersome input devices, specifying a query sketch may take undue amounts of time and learning the sketch conventions may discourage users from using such tools.

This report describes a computational framework and several algorithmic components towards an extensible solution to semantic event detection. The automated event detection algorithm enables users to effectively find semantically significant events in videos and help generate semantically meaningful highlights for fast browsing. In contrast to most existing event detection work, our goal is to develop an extensible computational approach for the detection of a range of events in different domains. We will demonstrate the proposed approach for the recognition of deciduous trees in still images and for the detection of hunt events in wildlife documentaries. The detection of deciduous trees can be achieved in a two layer architecture, where the first layer extracts texture and color measures and the second layer combines the measures to yield image labels indicating the presence or absence of deciduous trees at each image region. For the detection of hunts in video footage, we use a three-level algorithm. The first level extracts color, texture, and motion features, and detects moving object blobs. The mid-level employs a neural network to verify whether the moving blobs belong to objects of interest. This level also generates shot descriptors that combine features from the first level and contain results of mid-level, domain specific inferences made on the basis of shot features. The shot descriptors are then used by a domain-specific inference process at the third level to detect the video segments that contain events of interest. To test the effectiveness of our algorithm, we have applied it to

detect animal hunt events in wildlife documentaries. In our implementation we do not attempt to detect the stalking phase that precedes many hunts. Our purpose is to detect the swift or rapid chase of a fleeing or running animal. Since hunts are among the most interesting events in a wildlife program, the detected hunt segments can be composed into a program highlight sequence. The proposed approach can be applied to different domains by adapting the mid and high-level inference processes while directly utilizing the results from the low-level feature extraction processes.

In the following section, we describe the proposed computational framework and its algorithmic components. In Section 3, we present experimental results obtained as we applied the proposed algorithm to detection of animal hunt events in a number of commercially available wildlife video tapes. Implementation details are also furnished in Section 3. Finally in Section 4, we summarize our work and discuss some future directions.

II. Methodology

We focus on the classification and detection of non-rigid, amorphous or articulate natural objects, such as animals, trees, grass, sky, clouds, etc., as well as the motion of objects in such scenes. Our approach therefore has object classification and motion detection components. The object classification component makes use of feature extraction methods based on multi-resolution Gabor filters, the Gray-Level Co-occurrence Matrix (GLCM), the fractal dimension, and color. The feature representations of the objects are then classified by a back-propagation neural network. This concludes the task for object recognition in still images. For event detection in video data the classification labels are combined with shot boundary information and frame motion estimates to detect semantic events such as predators hunting prey.

The problem of detecting semantic events in video, e.g., hunts in wildlife video, can be seen as having three levels as shown in Figure 1. At the lowest level we determine the boundaries between shots, estimate the global motion, and express each frame in a color and texture space. We also compensate for the estimated global motion between each pair of frames. The earlier frame of each pair is transformed by the motion estimate and a difference image is produced to highlight areas of high residual error.We assume that this residual error is mostly due to independent object motion, and therefore the highlighted areas correspond to independently moving objects which are also referred as motion blobs (see Section III Figure 12).

At the intermediate level the detected motion blobs are then verified with the class labels

assigned to that region by a neural network. The network uses the color and texture representation of the input obtained by the lower level, and performs a crude classification of image regions into sky, grass, tree, rock, and animal regions. If (1) the motion between two consecutive frames is large, (2) a blob exists that has a high motion residual (motion other than that of the background), and whose motion and position in consecutive frames varies smoothly, and (3) is labeled as animal region by the network, then we assert that we are tracking a fast moving animal. The intermediate level generates and integrates such frame information and produces a summary for an entire shot. If throughout the shot there was support for a fast moving animal and the location/motion of the animal was found to be stable enough then the shot summary will indicate that a fast moving animal was tracked throughout the shot.

At the highest level a domain specific analysis of these shot summaries is used to infer the presence of a hunt in the underlying video sequence.

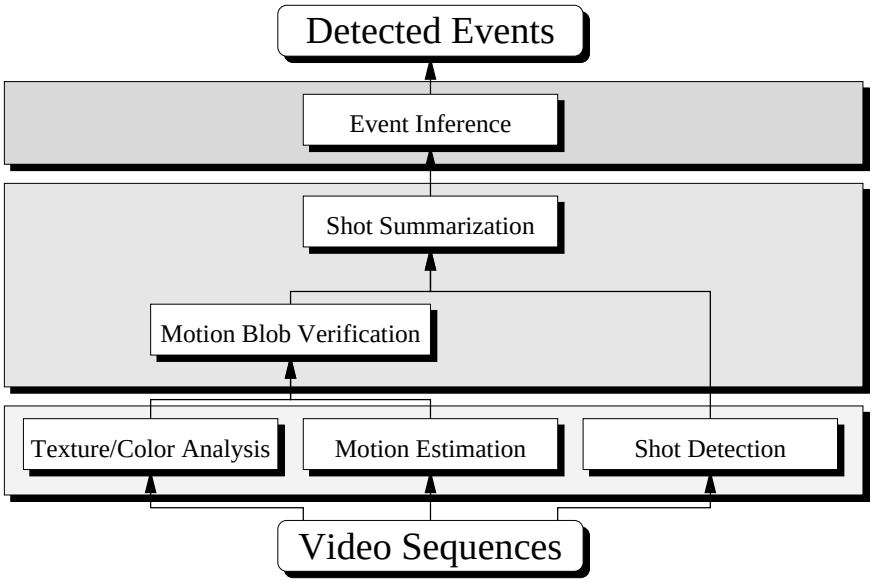


Fig. 1. The flowchart of our method. For object recognition in still images the method ends after the second level by providing object labels for each image region

A. Global Motion Estimation and Motion Blob Detection

We assume that the global motion can be estimated with a three parameter system allowing only for zoom, horizontal and vertical translation.

$$u(x, y) = a_0 + a_2 x$$

$$v(x, y) = a_1 + a_2 y$$

The robust recovery of the three parameters has to deal with the following problems,

- corresponding points in adjacent frames are often far apart (50-60 pixel displacements are not uncommon, peak displacements exceed 100 pixels),
- interlacing between frames drastically changes the appearance of small objects and textures in adjacent frames,
- the object and hence the global motion we are trying to estimate is often very large and motion blur eliminates texture in the direction of that motion (of course the motion in this direction is also the motion we are most interested in),
- often animals need to be tracked under strongly varying lighting conditions and occlusion, as when a hunt leads through trees or bushes.

Given the large possible displacements between corresponding patches of adjacent frames an exhaustive search of possible match locations creates unreasonable processing requirements. Therefore we use a pyramid of reduced resolution representations of each frame. At each level of the 5-level pyramid we consider matches from a 5×5 neighborhood around the location of the patch in the source frame, enabling a maximum matching distance of 62 pixels. The levels of the pyramid are obtained by subsampling the lower level image rather than computing a more accurate Gaussian pyramid. We expect the use of a Gaussian pyramid to produce better results at a slight computational cost.

At the lowest level of the pyramid, i.e. the full resolution representation of the frame, the patches used for matching are of size 64×64. Patches from uniform areas often result in erroneous displacement estimates. To avoid matching such patches we discard patches with insufficient "texture". We use a 2D variance measure to determine the "amount of texture".

$$\begin{aligned} var_x &= \sum_{y=0}^{n} (\sum_{x=0}^{m} (p(x,y) - p(.,y))^2 - q_x)^2 \\ var_y &= \sum_{x=0}^{m} (\sum_{y=0}^{n} (p(x,y) - p(x,.))^2 - q_y)^2 \end{aligned}$$

where p is an $m \times n$ image patch, $p(x,.)$ and $p(.,y)$ are the means of the x^{th} column and y^{th} row of p, and q_x and q_y are the means of $((p(x,y) - p(x,.))^2$ and $(p(x,y) - p(.,y))^2$ for all x and y within p, respectively.

We compute motion estimates at each of the four corners of a frame, as shown in Figure 11(a). Bad motion estimates are often due to matching errors made high up in the pyramid that are subsequently not recovered by the lower levels. Since the motion of the tracked animals often

does not vary drastically between consecutive frames (i.e. their acceleration is small) we also use the previous best motion estimate to predict the location of the four patches in the next frame. A limited search in a 5×5 neighborhood around the predicted location, improves the motion estimates in many cases. Therefore we obtain up to eight motion estimates, one pyramid based estimate for each of the four patch locations, and one for each of the four estimates based on a limited search around the predicted match locations. Since some patches may not pass the "texture" test we may have fewer than eight motion estimates. The highest normalized dot product between a source patch $P1$ and matched patch $P2$ determines the "correct" global motion estimate between the current and next frame. The normalized dot product is equal to the cosine of the angle (α) between the two patches (vectors) $P1$, and $P2$:

$$cos(\alpha)_{P1,P2} = \frac{\sum_{i,j} P1(i,j)P2(i,j)}{\sum_{i,j} P1(i,j) \sum_{i,j} P2(i,j)}$$

With respect to our particular task of detecting hunts in wildlife documentaries we would like to point out that

- almost all wildlife videos are taken with a tele lens at a great distance to the objects of interest. For our motion analysis, we therefore assume an orthographic model, in which the camera pan and tilt appear as plain translations, thus supporting our assumption of uniform background motion,
- motion estimates based on the *feature space representation* of the frames are very similar to those obtained on the original *color* frames, and
- although the described motion estimation scheme is sufficient for our purpose a Kalman filter based approach might yield more consistent results [4], [16].
- Alternative camera motion estimation schemes like Video Tomography methods [1] achieve similar results. Since they consider projections of entire frames they can get confused by moving objects.

The motion estimates are then used to compensate for the global motion between consecutive frames. Finally, we use the grayvalue difference between the current image and the motion compensated next frame to estimate the location of the animal in the frame. Areas with low residual error are assumed to have motion values similar to those of the background and are ignored. The independent motion of animals on the other hand usually causes high residual errors between the current frame and the following motion compensated frame. Therefore we can make use of a robust estimation technique to obtain an estimate of the animal location within

the frame. This estimation technique iteratively refines the mean x and y values dependent on the residual error within a fixed size neighborhood around the mean values for the entire difference image. The robust estimation method was first developed in [37] for real-time human face tracking. Here we briefly describe how the method is applied to the application discussed in this paper. Based on the frame difference result, the algorithm constructs two 1D histograms by projecting the frame difference map along its x and y direction, respectively. The histograms, therefore, represent the spatial distributions of the motion pixels along the corresponding axes. Figure 2(a) illustrates an ideal frame difference map where there is only one textured elliptical moving object in the input sequence, and the corresponding projection histograms.

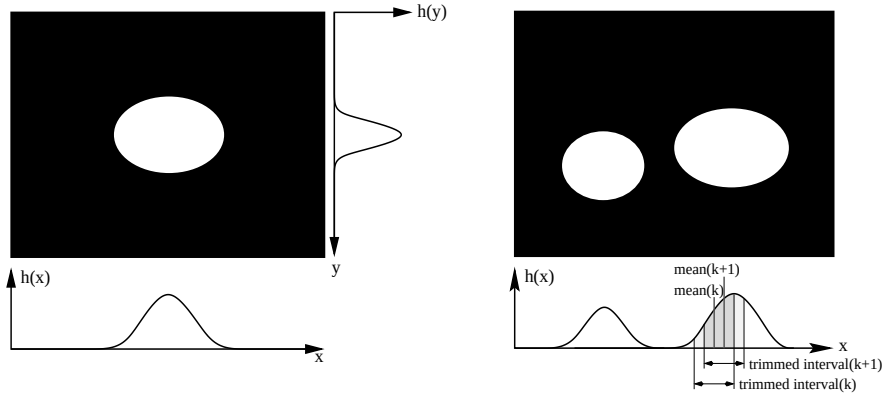


Fig. 2. (a) Two 1D histograms constructed by projecting the frame difference map along the x and y direction, respectively. (b) Robust mean estimation for locating the center position of a *dominant* moving object.

The instantaneous center position and size of a object in the image can be estimated based on statistical measurements derived from the two 1D projection histograms. For example, a simple method estimates the center position and size of a dominant moving object in an input sequence using the sample means and standard deviations of the distributions. More specifically, let $h_x(i), i = 0, 1, \ldots,$ and $h_y(i), i = 0, 1, \ldots,$ denote the elements in the projection histograms along the x and y direction, respectively. Then the object center position (x_c, y_c) and object width and height (w, h) may be estimated as:

$$x_c = \frac{\sum_i x_i h_x(i)}{\sum_i h_x(i)}, \quad y_c = \frac{\sum_i y_i h_y(i)}{\sum_i h_y(i)}, \quad w = \alpha \left[\frac{\sum_i (x_i - \mu_x)^2 h_x(i)}{\sum_i h_x(i)} \right]^{\frac{1}{2}}, \quad h = \beta \left[\frac{\sum_i (y_i - \mu_y)^2 h_y(i)}{\sum_i h_y(i)} \right]^{\frac{1}{2}}$$

where α and β are constant scaling factors.

However, the object center position and size derived from the sample means and standard deviations may be biased in the cases where other moving objects appear in the scene. It is therefore necessary to develop a more robust procedure to address this problem. We propose the use of robust statistical estimation routines to achieve robust measurements for object center position and size [46]. More specifically, the center position of a dominant moving object in an input sequence is estimated based on the robust (trimmed) means of the two 1D projection histograms in the x and y directions. Figure 2(b) illustrates the process of the estimation of the motion center.

Step 1 Compute sample mean μ and standard deviation σ based on all the samples of the distribution.

Step 2 Let $\mu_t(0) = \mu$ and $\delta = max(a\ \sigma, b{*}sampleSpaceWidth)$ where a and b are scaling factors, e.g., $a = 1.0$ and $b = 0.2$, and $sampleSpaceWidth$ is the image-width and image-height in the x and y direction, respectively.

Step 3 Compute trimmed mean $\mu_t(k+1)$ based on the samples within the interval $[\mu_t(k) - \delta, \mu_t(k) + \delta]$.

Step 4 Repeat Step 3 until $|\mu_t(k+1) - \mu_t(k)| < \epsilon$ where ϵ is the tolerance, e.g., $\epsilon = 1.0$. Denote the converged mean as μ^*.

Step 5 Let center-position $= \mu^*$.

In addition to the robust estimation of object center position, we propose the following routine for robust estimation of object size. The method first re-projects the frame difference result in a neighborhood of the located center. It then derives the object size based on the robust (trimmed) standard deviation. Given the robust mean μ^* and δ obtained from the above center locating routine, the routine for estimation the size in either x or y direction is as follows.

Step 1 Construct a clipped projection histogram H^{clip} by projecting the color filtering map within the range $[\mu^*_{opp} - \Delta, \mu^*_{opp} + \Delta]$ in the opposite direction, where μ^*_{opp} is the robust mean in the opposite direction and Δ determines the number of samples used in the calculation.

Step 2 Based on H^{clip}, compute the trimmed standard deviation δ_t based on the samples within the interval $[\mu^* - \delta, \mu^* + \delta]$.

Step 3 IF $H^{clip}(\mu^* + d\delta_t) \geq g\ H^{clip}(\mu^*)$ OR $H^{clip}(\mu^* - d\delta_t) \geq g\ H^{clip}(\mu^*)$,
where e.g., $d = 1.0$ and $g = 0.4$, THEN increase δ_t until the condition is no longer true.

Step 4 Let $size = c\ \delta_t$ where c is a scaling factor, e.g., $c = 2.0$.

B. Texture, Color and Motion Analysis: Low-Level Descriptors

To obtain rich, and hence robust and expressive descriptions of the objects in the video frames we describe each pixel in terms of color and texture measures. The color measures are the normalized red, green, and blue intensities of the pixel, and its grayvalue, while the texture measures are derived from the Gray Level Co-occurrence Matrix (GLCM), Fractal Dimension estimation methods, and a Gabor filter bank. The feature space representations of each pixel are classified into the categories sky/clouds, grass, trees, animal, rock using a back-propagation neural network. The use of these features in conjunction with the back-propagation classifier have previously been shown to enable the detection of deciduous trees in unconstrained images [20].

The rich image descriptions are formed from 56 Gray-Level Co-occurrence Matrix, 4 fractal dimension, 12 Gabor, and 4 color based measures. No one of the types of measure (e.g. color or Gabor measures) has the power of the combined set of measures. The neural network described in Section II-C is well suited to combine this set of measures and robustly classify image regions into various animal and non-animal classes. Note that we are only computing features from still frames and that motion is included explicitly at a higher level. In an alternative approach [40] uses temporal textures for classification, by combining spatial and temporal changes in image sequences.

B.1 Fourier Transform Measures

Some measures commonly used with Fourier Based Methods are i) wedge sampling, ii) annular-ring sampling, and iii) parallel-slit sampling.

Many textures differ significantly in the domains of the annular-ring and parallel-slit measures, however for our purpose of discriminating tree and non-tree areas angular wedge sampling is most expressive. Fourier Transforms (FTs) of images and image patches containing human-made structures often have line or wedge shaped areas of high spectral power that pass through the center as shown in the image/FT pair in Figures 3 (a) and (b). Summing the power in fixed angular intervals for all directions in the FT of the image lets us separate common from uncommon orientations in the image.

The shaded wedge in Figure 2 shows such an angular interval. A circular mask has been imposed so that the power in the diagonal directions is not unfairly biased. Once the power

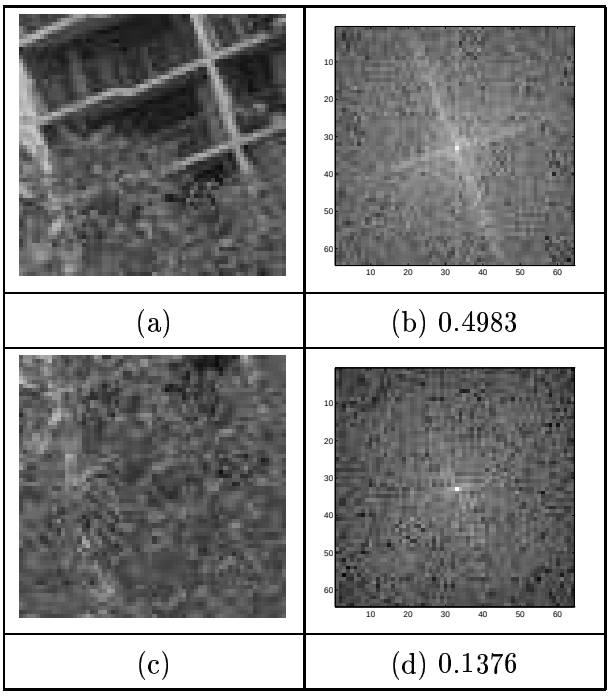


Fig. 3. An image containing human-made and tree areas (a) and its Fourier Transform (b). An image of leaves of a tree (c) and its Fourier Transform (d). The numbers associated with (b) and (d) are the structure measure (described in Section II-B.1) for images (a) and (c).

in each angular interval has been determined, we obtain the minimum and maximum angular power and use the normalized ratio $\frac{max-min}{max+min}$ to determine the amount of structure in the patch.

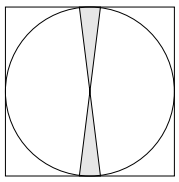


Fig. 4. The sum of the power of the Fourier Transform inside the shaded vertical angular interval is a measure of the "structure" present in an image patch.

Larger values for this wedge measure indicate greater "regularity" in some direction in the image patch, smaller values indicate less "regularity", in terms of parallel lines, bars and edges. Since we are comparing the ratio between the maximum and minimum value, this measure is rotation invariant.

Performing the above procedure on fixed-size image patches, we obtain local measures of the regularity of these patches. We obtained very similar results for patch sizes spanning three orders

of magnitude 16×16, 32×32, and 64×64 pixels.

B.2 Gabor Filter Measures

The image (in the spatial domain) is described by its 2-D intensity function. The Fourier Transform of an image represents the same image in terms of the coefficients of sine and cosine basis functions at a range of frequencies and orientations. Similarly, the image can be expressed in terms of coefficients of other basis functions. Gabor [19] used a combined representation of space and frequency to express signals in terms of "Gabor" functions:

$$F_{\theta,\nu}(\mathbf{x}) = \sum_{i=1}^{n} a_i(\mathbf{x})\, g_i(\theta, \nu) \tag{1}$$

where θ represents the orientation and ν the frequency of the complex Gabor function:

$$g_i(\theta, \nu) = e^{i\nu(xcos(\theta)+ysin(\theta))} e^{-\frac{x^2+y^2}{\sigma^2}} \tag{2}$$

Gabor filters have gained popularity in multi-resolution image analysis [17], [19], despite the fact that they do not form an orthogonal basis set. Gabor filter based wavelets have recently been shown [29] to be fast and useful for the retrieval of image data.

We convolve each image with Gabor filters tuned to four different orientations at 3 different scales. The average and range of the four measures at each scale are computed. To make the measurements somewhat scale invariant, we obtain the following four texture measures:

- The average of the orientation responses at all scales.
- The average of the scales' orientation response range.
- The range of the scales' averaged orientation responses.
- The range of the scales' orientation response range.

B.3 Steerable Filter Measures

Since many human-made structures exhibit a large amount of regularity in the form of parallel lines and bars, patches with few dominant orientations are less likely to represent trees. On the other hand, the irregular leaf and branch structure of trees often exhibits a greater variety of weak orientations.

Binning orientations appropriately, we can use the *number* and *strength* of different orientations in an image patch to distinguish between patches belonging to human-made scenes (which usually have fewer but stronger distinct orientations) and natural scenes.

Steerable bar and step edge detecting filters are used to obtain the dominant orientation for each image patch. The result of this routine is an orientation image indicating the orientation of the predominant step or bar edge at each location.

Perona [35] demonstrated a general constructive method to construct basis and interpolating functions and showed that all functions that are polar-separable with sinusoidal θ components are steerable. Examples of such functions are shown in Figure 5.

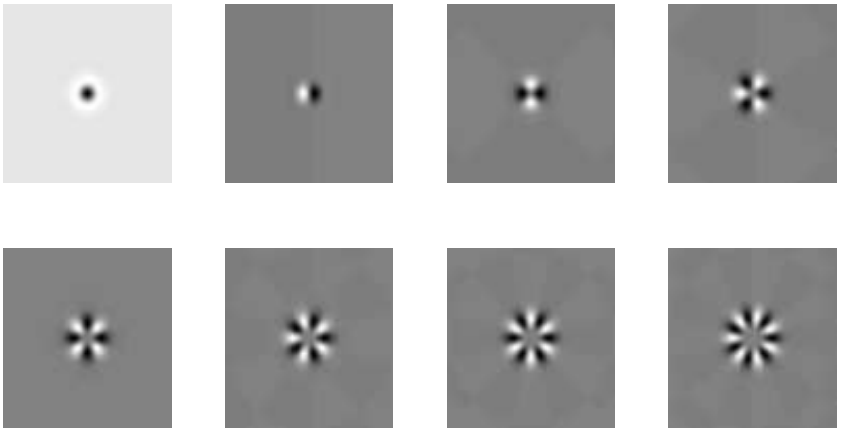


Fig. 5. Examples of polar separable functions with sinusoidal θ component corresponding to $a_0, \ldots, a_7$.

We used this method to obtain a steerable function set for a quadrature pair (G_{yy}, H_{yy}), where G_{yy} is the second derivative along the y-axis of an elongated Gaussian kernel $G(x, y, \sigma_x, \sigma_y) = e^{-((x/\sigma_x)^2+(y/\sigma_y)^2)}$ shown in Figure 6 (a) and H_{yy} is the Hilbert transform of G_{yy} shown in Figure 6 (b).

For multiple occurrences of lines and step edges, good angular resolution (orientation selectivity) was obtained when the ratio $\frac{\sigma_x}{\sigma_y}$ was at least $\frac{1}{4}$. Perona [35] shows an efficient method that places the second derivative of the Gaussian in the real part of the complex kernel and its Hilbert transform in the imaginary part.

The n-term approximation of the function we want to steer can be written as:

$$F_\theta^{[n]} = \sum_{i=1}^{n} \sigma_i \, a_i(\mathbf{x}) \, b_i(\theta) \tag{3}$$

where the σ_i weight the product of the i^{th} filter basis function a_i (the coefficients of the 2D Fourier series) and the corresponding interpolating function b_i (note that the b_i are the frequency basis functions of the Fourier Series).

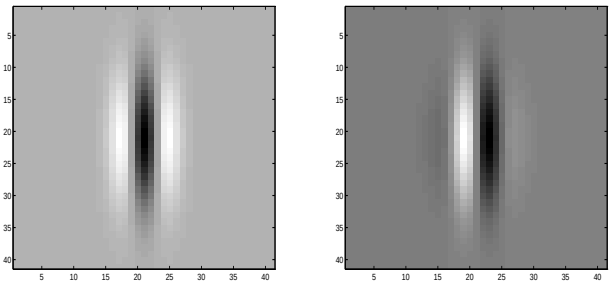


Fig. 6. Filters used to measure the energy in the image. The second derivative of an elongated Gaussian (left) is used to detect lines in an image. Its Hilbert transform (right) is used to detect step edges in an image.

The values for the σ_i, a_i and b_i are obtained by finding the Fourier series of the function $h(\theta)$, which is the integral of the product of the function with rotated versions of itself:

$$h(\theta) = \int_{I\!R^2} F_\theta(x)\overline{F_{\theta'=0}(x)}dx \tag{4}$$

where the integral ranges over all 2D space ($I\!R^2$) and $\overline{(.)}$ represents the complex conjugate. Note that $F_{\theta'=0}(x) = F(x)$.

Expanding $h(\theta)$ as a Fourier series we can read off the filter's (2D) basis functions a_i and the corresponding interpolating functions b_i.

$$\sigma_i = \sqrt{h(\nu_i)} \tag{5}$$

$$b_i(\theta) = e^{i\nu\theta} \tag{6}$$

$$a_i(x) = \sigma_i^{-1} \int_{S^1} \overline{F_\theta(x)} e^{i\nu\theta dx} \tag{7}$$

The σ_i terms are used only for error analysis. For details see [35].

These filters are used to obtain the oriented energy of both step as well as bar edges. Although we initially envisaged them to aid the recognition of deciduous trees in winter, when their leaves are missing, the orientation analysis also turned out to be useful for the recognition of leaves and trees in summer.

B.4 Graylevel Co-occurrence Matrix Measures

Let $p(i,j,d,\theta) = \frac{P(i,j,d,\theta)}{R(d,\theta)}$ where $P(.)$ is the graylevel co-occurrence matrix of pixels separated by distance d in orientation θ and where $R(.)$ is a normalization constant that causes the entries of $P(.)$ to sum to one.

In texture classification, the following measures have been defined, see for example [9], [21]:

The **Angular Second Moment (E)** (also called the Energy) assigns larger numbers to textures whose co-occurrence matrix is sparse.

$$E(d,\theta) = \sum_{j=1}^{N_g}\sum_{i=1}^{N_g}[p(i,j,d,\theta)]^2$$

The **Difference Angular Second Moment (DASM)** assigns larger numbers to textures containing only a few graylevel patches. This and other features use $p_{x-y}(n,d,\theta) = \sum_{\substack{j=1 \\ |i-j|=n}}^{N_g}\sum_{i=1}^{N_g} p(i,j,d,\theta)$

$$DASM(d,\theta) = \sum_{n=0}^{N_g} p_{x-y}(n,d,\theta)^2$$

The **Contrast (Con)** is the moment of inertia around the co-occurrence matrix's main diagonal. It is a measure of the spread of the matrix values and indicates whether pixels vary smoothly in their local neighborhood.

$$Con(d,\theta) = \sum_{n=0}^{N_g-1} n^2 \left[\sum_{\substack{j=1 \\ |i-j|=n}}^{N_g}\sum_{i=1}^{N_g} p(i,j,d,\theta) \right]$$

The **Inverse Difference Moment (IDM)** measures the local homogeneity of a texture. It weighs the contribution of the co-occurrence matrix entries inversely proportional to their distance to the main diagonal.

$$IDM(d,\theta) = \sum_{i=1}^{N_g-1}\sum_{j=1}^{N_g-1} \frac{1}{1-(i-j)^2} p(i,j,d,\theta)$$

The **Mean (M)** is similar to the contrast measure above but weights the off-diagonal terms linearly with the distance from the main diagonal, rather than quadratically as for the Contrast.

$$M(d,\theta) = \sum_{n=0}^{N_g-1} n \left[\sum_{\substack{j=1 \\ |i-j|=n}}^{N_g}\sum_{i=1}^{N_g} p(i,j,d,\theta) \right]$$

Similar to the Angular Second Moment the **Entropy (H)** is large for textures that give rise to co-occurrence matrices whose sparse entries have strong support in the image. It is minimal for matrices whose entries are all equally large.

$$H(d,\theta) = -\sum_{j=1}^{N_g}\sum_{i=1}^{N_g} p(i,j,d,\theta)\log\left(p(i,j,d,\theta)\right)$$

Other measures are, **Sum Entropy (SH)**, which uses $p_{x+y}(n,d,\theta) = \sum_{\substack{j=1 \\ |i+j|=n}}^{N_g}\sum_{i=1}^{N_g} p(i,j,d,\theta)$

$$SH(d,\theta) = -\sum_{n=0}^{2*N_g-1} p_{x+y}(n,d,\theta)\log\left(p_{x+y}(n,d,\theta)\right)$$

Difference Entropy (DH)

$$DH(d,\theta) = -\sum_{n=0}^{N_g} p_{x-y}(n,d,\theta)\log\left(p_{x-y}(n,d,\theta)\right)$$

Difference Variance (DV)

$$DV = -\sum_{n=2}^{2N_g}(n-DH)^2 p_{x-y}(n,d,\theta)$$

The **Correlation (Cor)** measure is an indication of the linearity of a texture. The degree to which rows and columns resemble each other strongly determines the value of this measure. This and the next two measures use $\mu_x = \sum_i i \sum_j p(i,j,d,\theta)$ and $\mu_y = \sum_j j \sum_i p(i,j,d,\theta)$.

$$Cor(d,\theta) = \frac{\sum_{i=1}^{N_g-1}\sum_{j=1}^{N_g-1} ijp(i,j,d,\theta) - \mu_x * \mu_y}{\sigma^2}$$

Shade (S) $$S(d,\theta) = \sum_i^{N_g}\sum_j^{N_g}(i+j-\mu_x-\mu_y)^3 p(i,j,d,\theta)$$

Prominence (P) $$P(d,\theta) = \sum_i^{N_g}\sum_j^{N_g}(i+j-\mu_x-\mu_y)^4 p(i,j,d,\theta)$$

Note that the directionality of a texture can be measured by comparing the values obtained for a number of the above measures as θ is changed. The above measures were computed at $\theta = \{0^o, 45^o, 90^o, \text{and } 135^o\}$ using $d = 1$. For further discussion of these graylevel co-occurrence matrix measures, see [9], [21], [45].

B.5 Fractal Dimension Measures

The underlying assumption for the use of the Fractal Dimension (FD) for texture classification and segmentation is that images or parts of images are self-similar at some scale.

Various methods that estimate the FD of an image have been suggested:

- Fourier-transform based methods [34],
- box-counting methods [8], [27], and
- 2D generalizations of Mandelbrot's methods [33].

The principle of self-similarity may be stated as: If a bounded set A (object) is composed of N_r non-overlapping copies of a set similar to A, but scaled down by a reduction factor r, then A

is self-similar. From this definition, the Fractal Dimension D is given by

$$D = \frac{log\, N_r}{log\, r}$$

The FD can be approximated by estimating N_r for various values of r and then determining the slope of the least-squares linear fit of $\frac{log\, N_r}{log\, \frac{1}{r}}$. The differential box-counting method outlined in Chaudhuri, *et al* [8] are used to achieve this task.

Three features are calculated based on

- the actual image patch $I(i,j)$,
- the high-graylevel transform of $I(i,j)$, $I_1(i,j) = \begin{cases} I(i,j) - L_1 & I(i,j) > L_1 \\ 0 & otherwise \end{cases}$
- the low-graylevel transform of $I(i,j)$, $I_2(i,j) = \begin{cases} 255 - L_2 & I(i,j) > 255 - L_2 \\ I(i,j) & otherwise \end{cases}$

where $L_1 = g_{min} + \frac{g_{avg}}{2}$, $L_2 = g_{max} - \frac{g_{avg}}{2}$, and g_{min}, g_{max}, and g_{avg} are the minimum, maximum and average grayvalues in the image patch, respectively.

The fourth feature is based on multi-fractals which are used for self-similar distributions exhibiting non-isotropic and inhomogeneous scaling properties. Let k and l be the minimum and maximum graylevel in an image patch centered at position (i,j), let $n_r(i,j) = l - k + 1$, and let $\mathcal{N}_r = \frac{n_r}{N_r}$, then the multi-fractal, D_2 is defined by

$$D_2 = \lim_{r \to 0} \frac{log \sum_{i,j} \mathcal{N}_r^2}{log\, r}$$

A number of different values for r are used and the linear regression of $\frac{log \sum_{i,j} \mathcal{N}_r^2}{log\, r}$ yields an estimate of D_2.

B.6 Entropy Measures

Since leaves and branches appear as rough and "messy" areas at most scales at which trees can be identified, we can use the entropy of image patches to separate them from uniform, smooth, and smoothly varying object surfaces. If V_{max} is the maximum value in an image patch, the entropy is defined as

$$Entropy = -\sum_{i=0}^{V_{max}} h_i\, log(h_i)$$

where $h_i = \frac{n_i}{N}$ is the i^{th} histogram count n_i divided by the total number of pixels in the image patch (N). We measure the entropy in both the gray value image as well as the orientation image described above, both measures are largely rotation invariant.

For motion analysis we use the histograms of the intensity values of patches in two consecutive frames as samples of their probability density functions and compute the **Mutual Information**,

$$H(X;Y) = H(X) - H(X|Y),$$

the **Entropy Distance**,

$$D_H(X,Y) = H(X,Y) - H(X;Y),$$

and the **Kullback-Leibler Divergence**,

$$D_{KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

between the distributions of the samples to determine whether a region corresponds is moving or static.

B.7 Color Measures

While the intensities of the red, green and blue components of a color image are highly correlated, the hue, saturation, and value decomposition offers a more independent representation that captures complementary information of the image.

We also use opponent color measures that contrast the intensities of Red vs. Green ($\frac{Red}{Green+\alpha}$), Red vs. Blue ($\frac{Red}{Blue+\alpha}$), and Green vs. Blue ($\frac{Green}{Blue+\alpha}$), where we used $\alpha = 0.01$ to bound the ratios.

C. Region Classification and Motion Blob Verification

We use a back-propagation neural network to arbitrate between the different features describing the image. Our back-propagation neural network [15] has a single hidden layer and uses the sigmoidal activation function $\Phi(act) = \frac{1}{1+e^{-act}} - 0.5$, where act is the activation of the unit before the activation function is applied. A single hidden layer in a back-propagation neural network has been shown to be sufficient to uniformly approximate any function (mapping) to arbitrary precision [11]. Although this existential proof doesn't state that the best network for some task has a single hidden layer, we found one hidden layer adequate. The architecture of the network is shown in Figure 7. The back-propagation algorithm propagates the (input) function values layer by layer, left to right (input to output) and back-propagates the errors layer by layer, right to left (output to input). As the errors are propagated back to the input units, part of each unit's error is being corrected.

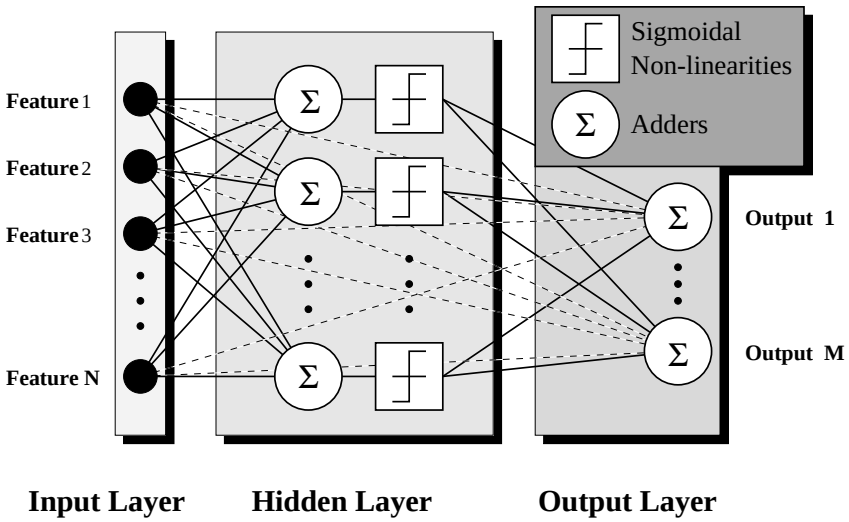


Fig. 7. The Neural Network architecture.

For the deciduous tree detection example we trained the network using only one label indicating whether a pixel is part of a tree and a non-tree image region.

For the hunt detection example we trained the network using a total of 14 labels. 9 animal labels (lion, cheetah, leopard, antelope, impala, zebra, gnu, elephant, and an all-other-animal class) and 5 non-animal labels (rock, sky/clouds, grass, trees, and an all-other-non-animal class) as well as a don't care label that was used to tell the network to ignore border regions between instances of the different groups, which arguably are bad training inputs.

After training, the tree detecting network produced the results shown in Figure 9 and 10.

For the hunt detection example we found that the proposed network performed well at classifying grass, trees, rocks, sky, and animals as a whole group. However, it is difficult for the network to classify lions, cheetahs, leopards, antelopes, impalas, gnus, hyenas, and even zebras, rhinos and elephants each into different groups. This is probably due to the fact that those animals differ mostly in their shape and size which we do not model. Hence, while the network was still trained on the different animal labels, we artificially grouped those labels into a single "animal" label when using the network for animal region verification. We also found that the network did not perform well at solving the opposite problem of classifying, grass, trees, rocks, and sky together as a single "non-animal" group. The differences between the appearance of instances of these groups are severe. Asking the network to assign one label to them and a different label to animals proves to be more difficult than the classification into the individual non-animal groups.

The output of the animal detecting network is then used to verify the motion blob candidates from section II-A. In our current implementation, a simple procedure is employed which implements the following test. A region that has high residual motion after motion compensation and that contains a significant amount of animal labels, as detected by the neural network, is considered as a possible moving animal region.

D. Shot Summarization and Intermediate-Level Descriptors

We use a simple color histogram based technique to decompose video sequences into shots. To avoid missing important events in extended shots, we also *force* a shot summary every 200 frames. A third kind of shot boundary is inserted whenever the direction of the global motion changes. Shot boundaries of this last kind ensure that the motion within shots is homogeneous. Each shot is then summarized in terms of intermediate-level descriptors. The purpose of generating intermediate-level shot summaries is two-fold. First, the shot summaries provide a way to encapsulate the low-level feature and motion analysis details so that the high-level event inference module may be developed independent of those details, rendering it robust against implementational changes. Second, the shot summaries abstract the low-level analysis results so that they can be read and interpreted more easily by humans. This simplifies the algorithm development process and may also facilitate video indexing, retrieval and browsing in video database applications.

In general, the intermediate-level descriptors may consist of (1) *object*, (2) *spatial*, and (3) *temporal* descriptors. The *object* descriptors, e.g., "animal", "tree", "sky/cloud", "grass", "rock", etc. indicate the existence of objects in the video frames. The *spatial* descriptors represent the location and size information about objects and the spatial relations between them in terms of spatial prepositions such as "inside", "next to", "on top of", etc. [12], [13]. The *temporal* descriptors represent motion information about objects and the temporal relations between them in terms of temporal prepositions such as "while", "before", "after", etc. [12], [13].

For the hunt detection application, we currently employ a particular set of intermediate-level descriptors which describe: (1) whether the shot summary is due to a forced or detected shot boundary; (2) the frame number of the beginning of the shot; (3) the frame number of the end of the shot; (4) the global motion; (5) the object motion; (6) the initial object location; (7) the final object location; (8) the initial object size; (9) the final object size; (10) the smoothness of the motion; (11) the precision throughout shot; and (12) the recall throughout shot. More

precisely, the motion descriptors provide information about the x- and y- translation and zoom components of motion. The location and size descriptors indicate the location and size of the detected dominant motion blob at the beginning and the end of the shot. The precision is the average ratio of the number of animal labels within the detected dominant motion blob versus the size of the blob, while the recall is an average of the ratio of the animal labels within the detected dominant motion blob versus the number of animal labels in the entire frame. In addition, we also employ descriptors indicating (13) that tracking is engaged; (14) that object motion is fast; (15) that an animal is present; (16) the beginning of a hunt; (17) number of consecutive hunt shot candidates found; (16) the end of a hunt; and (19) whether a valid hunt is found. See Section III-B.6 for an example and further explanation.

E. Event Inference

Hunt events are detected by an event inference module which utilizes domain-specific knowledge and operates at the shot level based on the generated shot summaries. From observation and experimentation with a number of wildlife documentaries, a set of rules have been deduced for detecting hunts. The rules reflect the fact that a hunt usually consists of a number of shots exhibiting smooth but fast animal motion which are followed by subsequent shots with slower or no animal motion. In other words, the event inference module looks for a prescribed number of shots in which (a) there is at least one animal of interest; (b) the animal is moving in a consistently fast manner for an extended period; and (c) the animal stops or slows down drastically after the fast motion. Figure 8 shows and describes a state diagram of our hunt detection inference model.

Automatic detection of the properties and sequences of actions in the state digram is non-trivial and the low-level feature and motion analysis described earlier in this paper are necessary to realize the inference. Since any event can be defined by the occurrence of objects involved and the specification of their spatio-temporal relationship, the proposed mechanism, of combining low-level visual analysis and high-level domain-specific rules, may be applicable to detect other events in different domains. In Section III-B.7, we provide an example and further explanation for using this inference model for hunt detection.

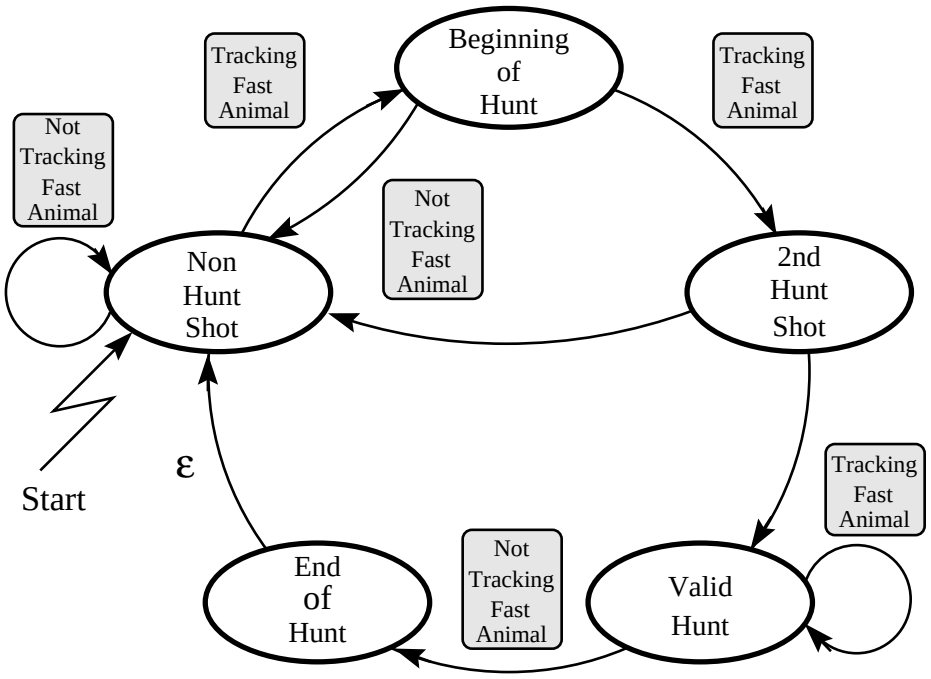


Fig. 8. The state diagram of our hunt detection method. Initially the control is in the Non-Hunt state on the left. When a fast moving animal is detected the control moves to the Beginning of Hunt state at the top of the diagram. When three consecutive shots are found to track fast moving animals then the Valid Hunt flag is set. The first shot afterwards that does not track a fast moving animal takes the control to the End of Hunt state, before again returning to the Non-Hunt state.

III. Results

This section describes the results for the object and event detection methods. The object recognition performance is illustrated for deciduous trees in unconstraint images. The event detection performance is illustrated for the detection of hunts in wildlife documentaries.

A. Object Recognition

A.1 The Performance of the Resulting Feature Set

Measures of every fifth pixel of 19 training images were obtained (well over half a million data points; the 51-D data set is about 119 MB; the 13-D data set is about 30 MB) and combined with labeled images to train the network. Subsampling speeds up the training process without (noticeably) affecting its outcome, since neighboring pixel locations are highly correlated.

We would like to point out that some of the test images in Figure 9 and Figure 10 show trees in fall with the leaves' colors ranging from green, through yellow, orange and red to a magenta-ish

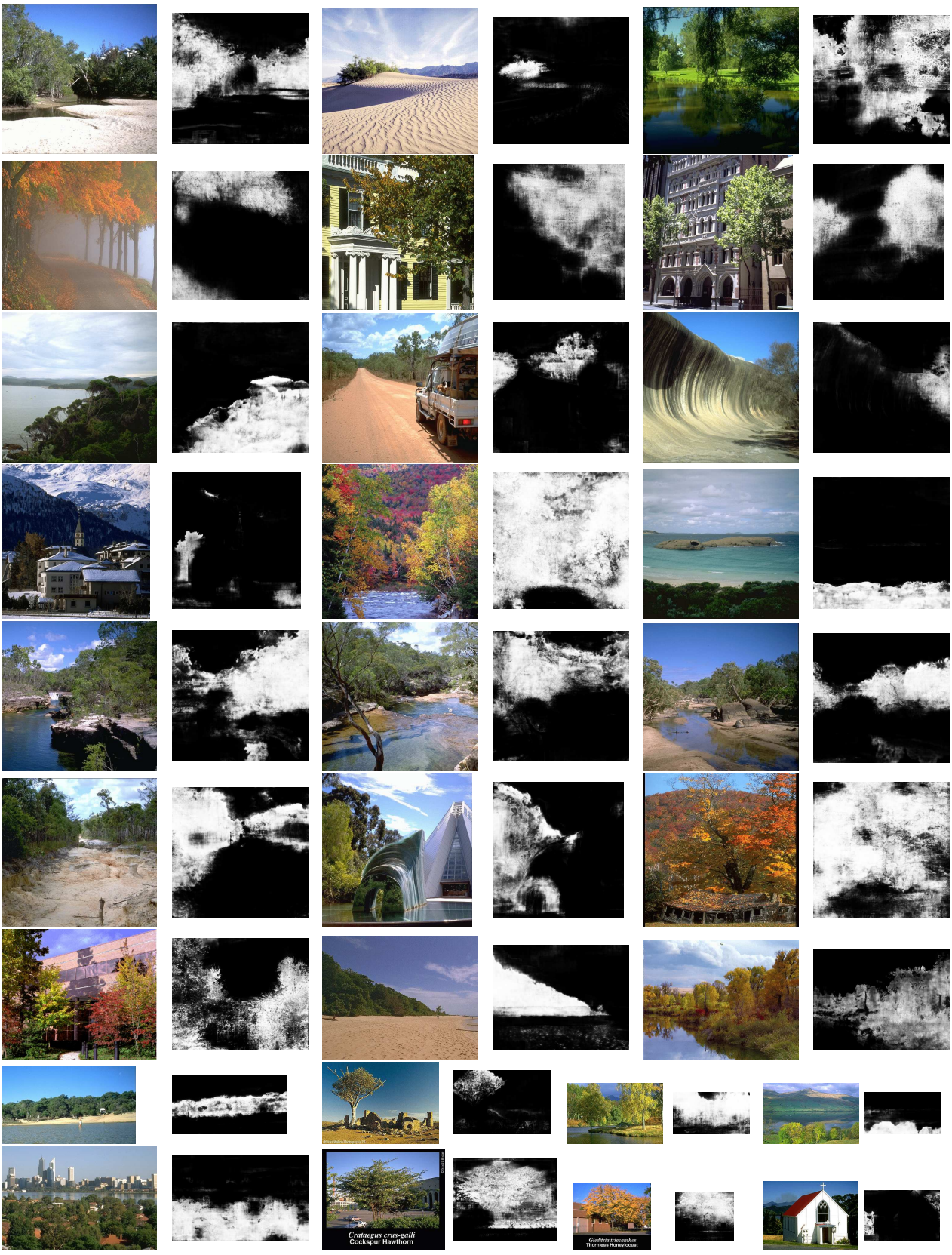


Fig. 9. Test images and the corresponding classification result s.

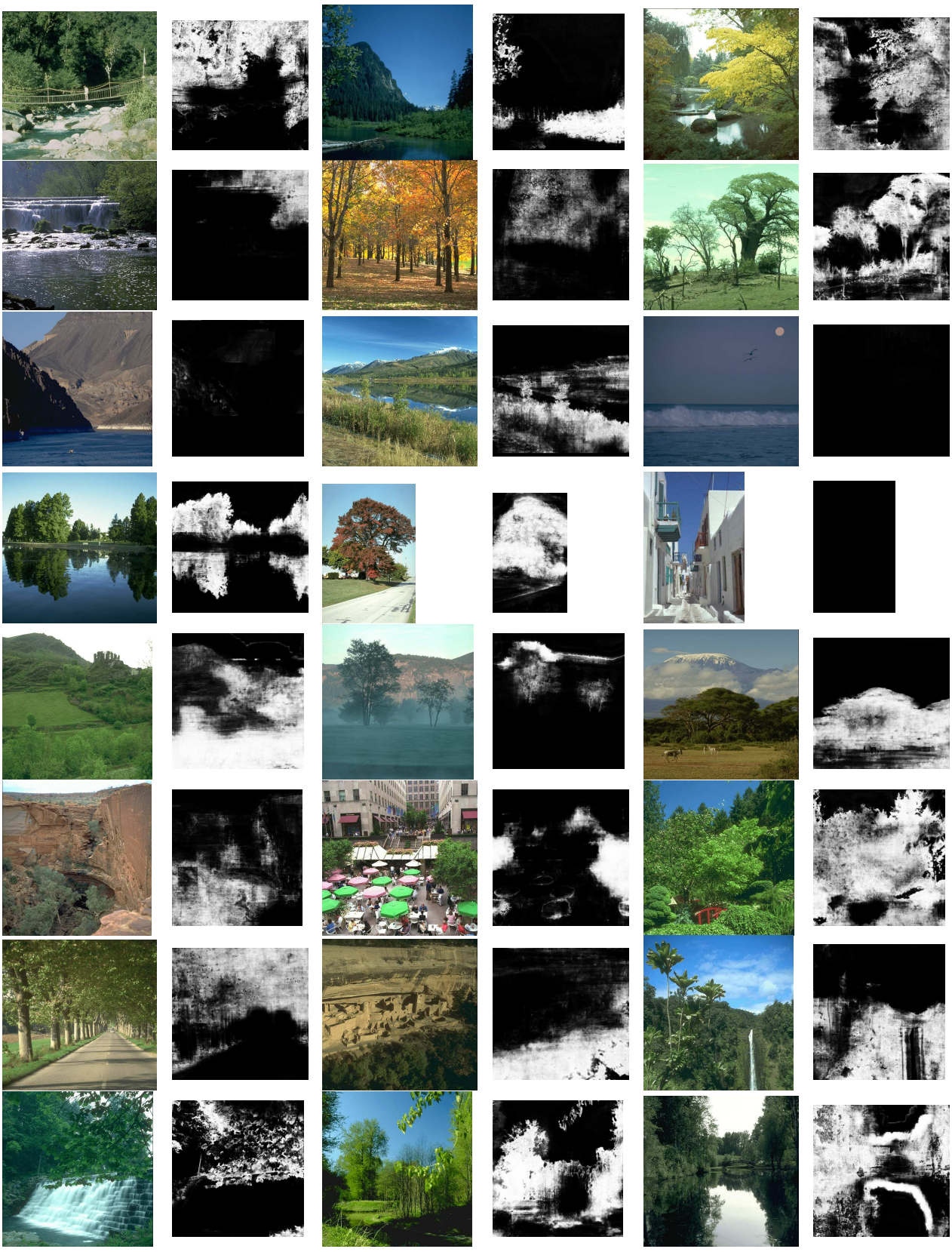


Fig. 10. Test images and the corresponding classification results.

red. Color is often a useful cue, but the network has also learned that leaves are not always green and not everything green depicts leaves. The same is true for the other features.

The first image on the second row in Figure 9 shows the performance of the approach for an image taken on a foggy day, with low contrast and low color saturation. The second image on the fourth row in Figure 9 shows the approach's robustness with respect to scale and color. This fall image shows trees at distances ranging between 5 meters and over 500 meters whose colors range from magenta to green. The output image shows that almost all tree regions were correctly labeled.

B. Event Recognition

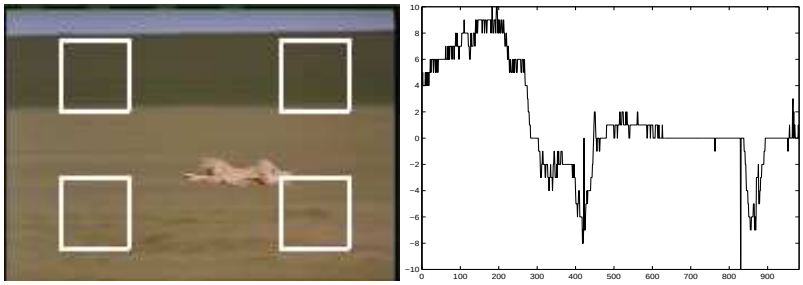
The proposed algorithm has been implemented in C++ and tested on Sun workstations. To evaluate the effectiveness of the algorithm, we have digitized wildlife video footage from a number of commercially available VHS tapes from different content providers. In the following sections we show examples of the extracted texture and color features, the motion estimation and detection results, the region classification results, the shot summaries, and the final hunt event detection results.

B.1 Test Data

About 45 minutes of actual wildlife video footage have been digitized and stored as test data for our hunt detection experiments. The frame rate of the video is 30 frames per second and the digitized frame resolution is 360 x 243 pixels. A total of 10 minutes of footage $\stackrel{\triangle}{=}$ 18000 frames $\stackrel{\triangle}{=}$ 100 shots have been processed so far.

B.2 Global Motion Estimation

Figure 11(a) shows the size and locations of the four regions at which the global motion is estimated. For each pair of frames motion estimates are computed using a 5 level pyramid scheme at the shown patch locations. In addition the previous motion estimate is taken as the current motion estimate and a tight local search around the four *predicted* patch locations yields another four patch matches. The best match of any of these 8 patch comparisons becomes the motion estimate for the current frame pair. Figure 11(b) shows the motion estimates during a hunt.



(a) (b)

Fig. 11. (a) The locations used to estimate the global motion, and (b) the motion estimates during a hunt.

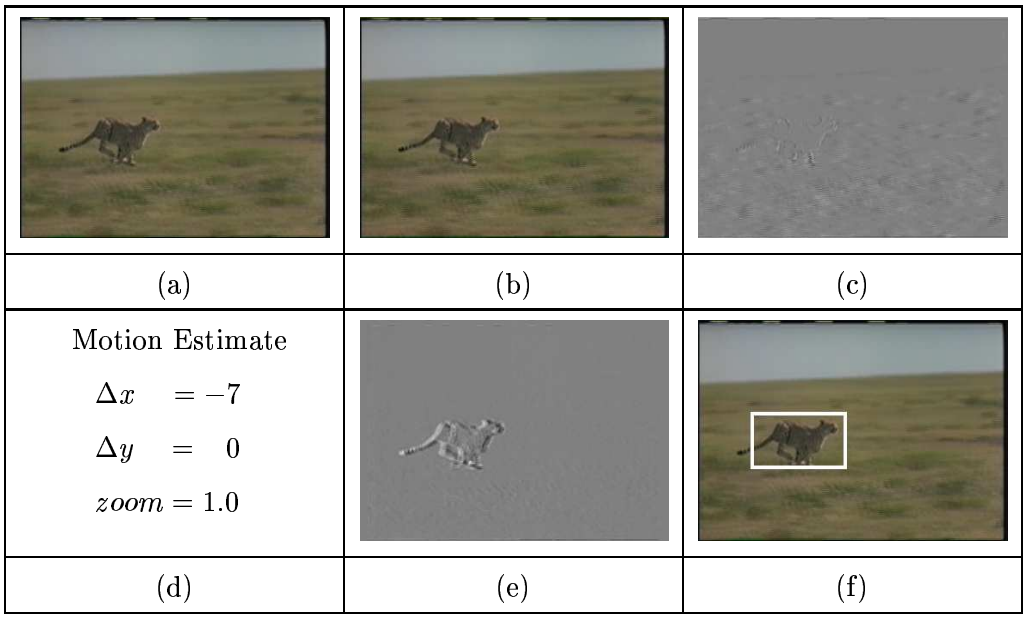


Fig. 12. Two consecutive frames from a hunt (a) and (b), the difference image (c), the estimated motion between the two frames (d), the motion compensated difference image (e), and the box around the area of largest residual error in the motion compensated difference image.

B.3 Motion Blob Detection

Figure 12 shows an example of the motion blob detection results. It is apparent that reliable estimation and compensation of global motion makes the task of motion blob detection relatively easier. When the accuracy of the global motion estimation results are poor, the performance of the motion blob detection relies largely on the robustness of the motion blob detection and tracking algorithm described in Section 2.1.

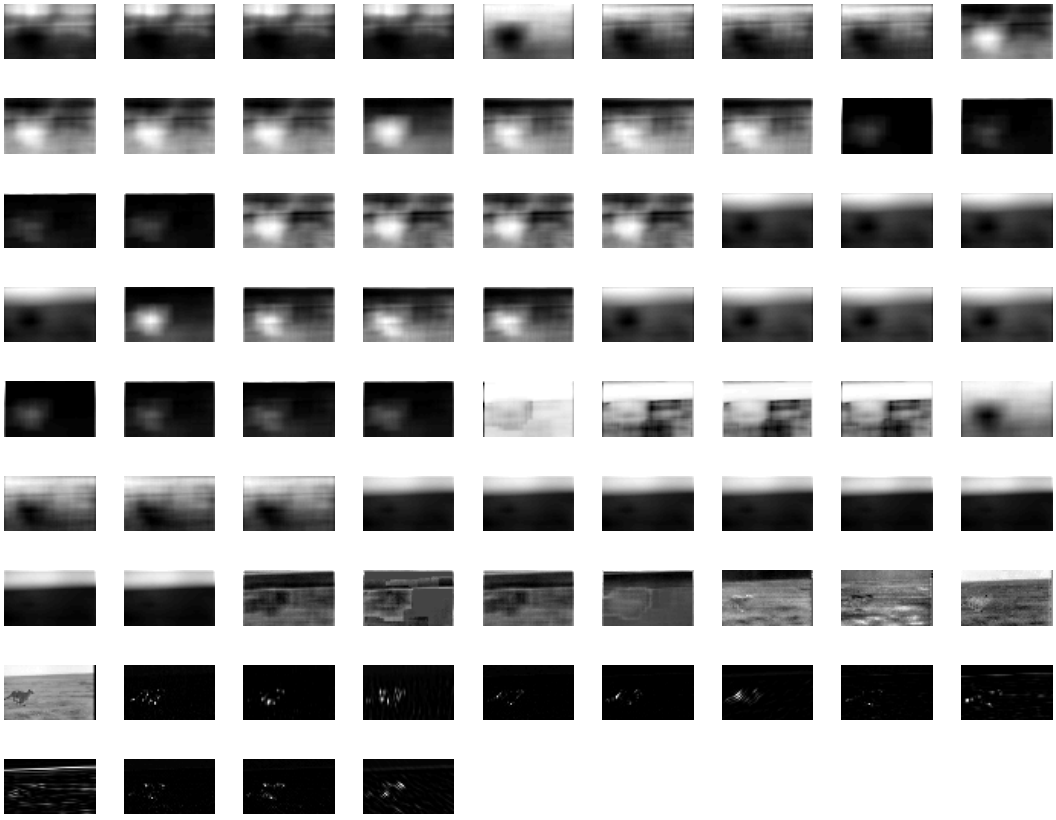


Fig. 13. The feature space representation of the first frame in Figure 12.

B.4 Feature Space Representation of the Video Frames

Figure 13 shows the feature space representation of a video frame. The features shown are the results of the Gray-Level Co-occurrence Matrix based measures (first 56 feature images), the Fractal Dimension based measures (next 4 feature images), the color based measures (next 4 feature images), and the Gabor based measures (last 12 feature images).

B.5 Region Classification

A neural network is trained on a number of training frames from wildlife video. The network is then used to classify unseen wildlife video. Global motion estimates such as the ones in Figure 11 are used to detect moving objects as shown in Figure 12. The locations of these moving object blobs are then verified using a neural network image region classifier that combines color and texture information. Rows 1, 3, and 5 of Figure 14 show a number of frames from hunts together with their classification results (rows 2, 4, and 6).

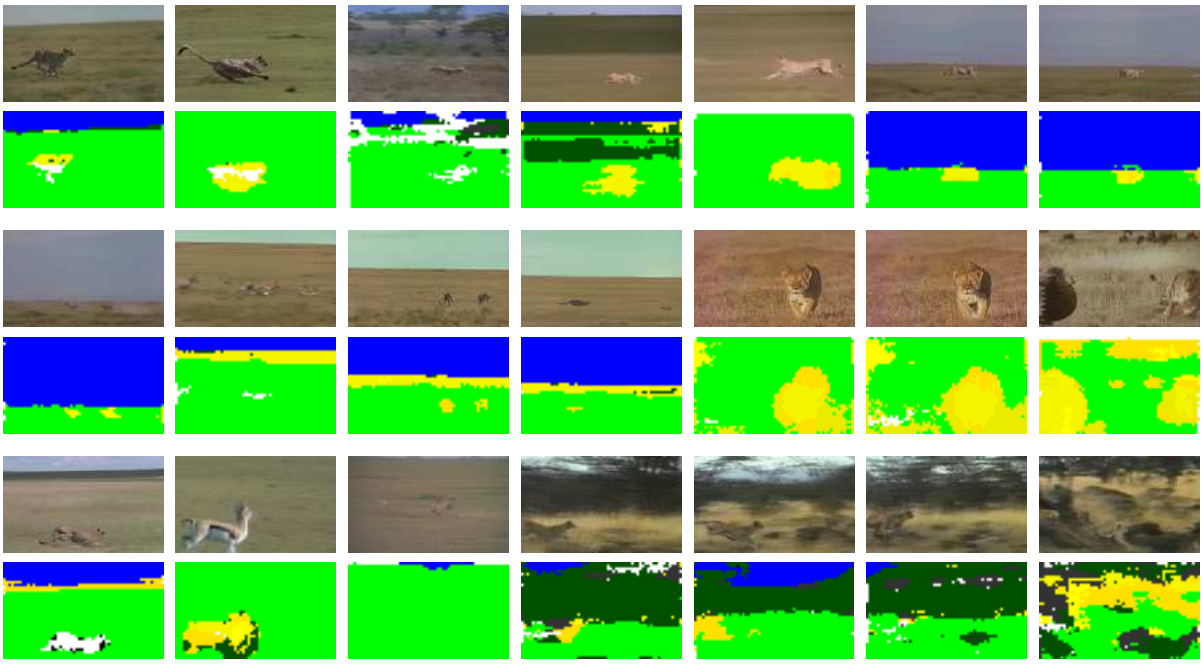


Fig. 14. Color and texture based segmentation results.

B.6 Shot Summarization

The intermediate level process consists of two stages. In the first stage the global motion estimates are analyzed and directional changes are detected in the x and y directions. When the *signs* of the 50 frame global motion averages before and after the current frame differ and their *magnitudes* are greater than 1 pixel per frame we insert an artificial shot boundary. In the second stage each shot is then summarized as in the example shown below.

```
---------- General Information ------------
Forced/real shot summary                  : 0
First frame of shot                       : 64
Last frame of shot                        : 263
Global motion estimate (x,y)              : (-4.48, 0.01)
Within frame animal motion estimate (x,y) : (-0.17, 0.23)
Initial position (x,y)                    : (175,157)
Final position (x,y)                      : (147,176)
Initial size (w,h)                        : ( 92, 67)
Final size (w,h)                          : (100, 67)

------- Hunt Information -------
Tracking                          : 1
Fast                              : 1
Animal                            : 1
Beginning of hunt                 : 1
Number of hunt shot candidates    : 1
End of hunt                       : 0
Valid hunt                        : 0
```

```
Motion smoothness throughout shot (x,y)   : ( 0.83, 0.75)
Precision throughout shot                 : ( 0.84)
Recall throughout shot                    : ( 0.16)
```

The summary consists of two parts, the first part, under `General Information` shows general statistics extracted for this shot, while the second, under `Hunt Information` consists of inferences based on those statistics for the hunt detection application.

The first row of the general Information part of the summary shows whether the shot boundary corresponding to this shot summary was real, i.e. whether it was detected by the shot boundary detector, or if it was forced because the maximum number of frames per shot was reached or the global motion has changed. The next two rows show the first and last frame numbers of this shot. The following measurements are shot statistics, i.e., the average global motion over the entire shot on row four, and the average object motion within the shot on row five. The next four rows measure the initial position and size, as well as the final position and size of the detected dominant motion blob. The third last row shows the smoothness of global motion where values near 1 indicate smooth motion and values near 0 indicate unstable motion estimation. The detection of a reversal of the global motion direction, described above, was based on a long term average of the motion estimates around the current frame, indicates a *qualitative* change in the global motion. The smoothness measure described here, on the other hand, provides a *quantitative* measure of the smoothness of the estimated motion. Finally the last two rows show the average precision and recall for the entire shot. As defined in Section II-D, the precision is the average ratio of the number of animal labels within the detected dominant motion blob versus the size of the blob, while the recall is an average of the ratio of the animal labels within the detected dominant motion blob versus the number of animal labels in the entire frame.

The hunt information part of the shot summary shows a number of predicates that were inferred from the statistics in part one. The shot summary shown above summarizes the first hunt shot following a `forced` shot boundary. The system is indicating that it is `Tracking` a `Fast` moving `Animal` and hence, that this could be the `Beginning of a hunt`. The `Tracking` predicate is true when the motion smoothness measure is greater than a prescribed value and the motion blob detection algorithm detects a dominant motion blob. The `Fast` predicate is set to true if the translational components of the estimated global motion are sufficiently large in magnitude, and the `Animal` predicate is true if the precision, i.e. the number of animal labels

within the tracked region, is sufficiently large. (The recall measure has not been used in our current implementation.) The remaining predicates are determined and used by the inference module as described below.

B.7 Event Inference and Final Detection Results

The event inference module infers the occurrence of a hunt based on the intermediate descriptors as described in Section III-B.6. In doing so, it employs four predicates, `Beginning of hunt`, `Number of hunt shot candidates`, `End of hunt`, and `Valid hunt`, which are currently embedded in the shot summary. If the intermediate descriptors `Tracking`, `Fast` and `Animal` are all true for a given shot, the inference module sets `Beginning of hunt` to be true, which means the shot could potentially be the beginning of a hunt event. The inference module tracks the intermediate descriptors `Tracking`, `Fast` and `Animal` for consecutive shots and increments the value of the `Number of hunt shot candidates` if all those three descriptors hold true for consecutive shots. In our current implementation, when the `Number of hunt shot candidates` is equal or greater than 3, `Valid hunt` is set to be true. Finally the inference module sets `End of hunt` to be true if one of the intermediate descriptors `Tracking`, `Fast` and `Animal` becomes false, which implies either the animal is no longer visible or trackable, or the global motion is slow enough indicating a sudden stop after fast chasing.

In our final results, hunt events are specified in terms of their starting and ending frame numbers. In the 10 minutes (18000 frames) of wildlife video footage which we have processed, there exist 7 hunt events. Table I shows the actual frames of the 7 hunts and all the frames of the detected hunts when we applied the proposed algorithm to the 10 minute video footage. The table also shows the retrieval performance of our method in terms of the two commonly used evaluation criteria (1) precision and (2) recall.

IV. Summary and Discussion

In this paper, we have presented a new computational framework and a number of enabling algorithmic components for automatic event detection in video and applied it to the detection of deciduous trees in still images and hunts in wildlife documentaries.

TABLE I

A COMPARISON OF THE ACTUAL AND DETECTED HUNTS IN TERMS OF THE FIRST AND LAST HUNT FRAME, AND THE ASSOCIATED PRECISION AND RECALL.

Sequence Name	Actual Hunt Frames	Detected Hunt Frames	Precision	Recall
hunt1	305 - 1375	305 - 1375	100 %	100 %
hunt2	2472 - 2696	2472 - 2695	100 %	99.6%
hunt3	3178 - 3893	3178 - 3856	100 %	94.8%
hunt4	6363 - 7106	6363 - 7082	100 %	96.8%
hunt5	9694 - 10303	9694 - 10302	100 %	99.8%
hunt6	12763 - 14178	12463 - 13389	67.7%	44.2%
hunt7	16581 - 17293	16816 - 17298	99.0%	67.0%
Average			95.3%	86.0%

A. Object Recognition

The use of a back-propagation network offers a simple solution to the laborious task of finding a good combination of the features. We have shown that feature sets like the one presented have sufficient expressive power to allow good generalization from only a few training images. Since the back-propagation algorithm is well understood and analyzed we have shown that it is possible to determine the usefulness of a specific feature if we had to reduce the amount of features used to an subset. The neural network approach offers a synthetic solution to the sensor fusion problem that is concerned with combinations of (possibly dependent) features for the purpose of classification and/or recognition. An analytical approach on the other hand would be difficult to conduct since the interactions even between modest numbers of dependent features are complex.

B. Event Detection

Our experimental results have verified the effectiveness of the proposed algorithm. The developed framework decomposes the task of extracting semantic events into three stages where visual information is analyzed and abstracted. The first stage extracts low-level features and is entirely domain-independent. The second stage analyzes the extracted low-level features and generates intermediate-level descriptors some of which may be domain-specific. In this stage,

shots are summarized in terms of both domain-independent and domain-specific descriptors. To generate the shot summaries, regions of interest are detected, verified and tracked. The third and final stage is domain-specific. Rules are deduced from specific domains and an inference model is built based on the established rules. In other words, each lower stage encapsulates low-level visual processing from the higher stages. Therefore the processes in the higher stages can be stable and relatively independent of any potential detail changes in the lower level modules. In order to detect different events, the expected changes are (a) the addition of descriptors in the second stage and (b) the design of a new set of rules in the third stage. The proposed algorithm also provides several reusable algorithmic components. In fact, the extracted low-level texture and color features are domain independent and many objects involved in events carry certain texture and color signatures. The neural network used for image region classification can be easily re-configured or extended to handle other types of objects [20]. The robust statistical estimation based object tracking method has already been used in different applications and its robustness and simplicity are verified in experiments repeatedly [37].

It is important for us to point out that the proposed algorithm detects hunt events by detecting spatial-temporal phenomena which are physically associated with a hunt event in the nature. More precisely, the physical phenomenon which we attempt to capture is the combination of the presence of animals in space and their movement patterns in time. This is in contrast to many existing event detection methods which detect events by detecting artificial postproduction editing patterns or other artifacts. The drawbacks of detecting specific editing patterns or other artifacts are that those patterns are often content provider dependent and it is difficult, if not impossible, to modify the detection methods and apply them to the detection of other events. It is also important to point out that our algorithm solves a practical problem and the solution is needed in the real world. In the wildlife video tapes which we obtained, the speech from the audio track and the text from the close-caption are loosely correlated with the visual footage. It is therefore unlikely that the hunt segments may be accurately located by analyzing the audio track and close-caption. In other words, given the existing wildlife tapes, a visual-information-based detection algorithm is needed to locate the hunt segments otherwise manual annotation is required.

V. Future Work

We propose to investigate the power and usefulness of temporal texture and color measures for the combined purpose of motion analysis, object and event detection.

A. Goals and Investigations

The goals of this investigation are

- the demonstration of the usefulness of *simultaneous spatio-temporal* analysis of video data,
- the derivation of *predicates* that are useful for the description of video data (objects, actions, events, etc.),
- methods and indicators that provide reliable *confidence values* for these predicates,
- the demonstration of the usefulness of the derived predicates in terms of an example representing an important class of problems.

B. Qualitative Motion Estimates

But, for many problems, it is possible to infer useful abstractions without full motion information for each pixel. Qualitative motion estimates, such as, the *onset* or *offset* of a motion, the *occlusion/disocclusion* of objects by other objects, the *tracking* of an object by the camera, the *egomotion* of an object, the *triggering* or *ending* of motion of one object by another object, are fundamental motion patterns that are likely to be useful for a range of higher level video analysis tasks.

Figure 15 shows models for these fundamental motion patterns given small camera motion or a stationary camera.

For a valid (dis)occlusion regions 1 and 3 (before and after the (dis)occlusion) must have identical or very similar texture and color characteristics, while for region 2 the characteristics may be different.

Note that the small camera motion condition is not a restriction, but merely a fact that we need to verify or reject. Once we have established that the camera is moving, we can try to determine whether it is tracking something. If on the other hand we found that the camera is stationary, we can try to locate moving objects. But it is important to state the camera motion condition for the definition of the above qualitative motion primitives.

Either way we must make the assumption that moving objects occupy small areas in static scenes. If this is not true, we need to recognize the objects first and then use background

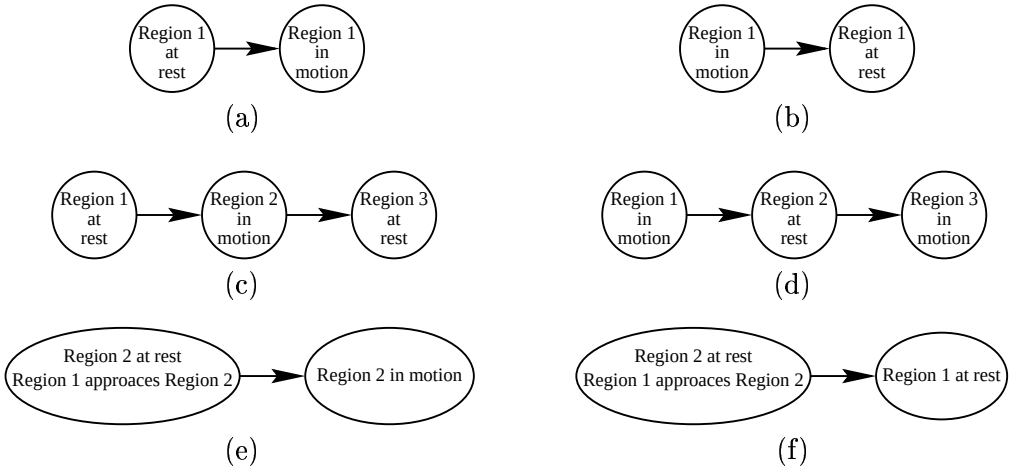


Fig. 15. The transition diagrams for stationary cameras. (a) shows the motion onset model, (b) shows the motion offset model, (c) shows the temporal occlusion model, (d) shows the temporal disocclusion model, (e) shows the trigger-motion model, and (f) shows the end-motion model.

knowledge to determine absolute motion.

B.1 Tracking and Motion Detection

In Section II-A we assumed that that the moving objets in a video occupy only small areas. In this case we establish that we are **tracking** an object when regions that exhibit rest characteristics are surrounded by regions that exhibit motion characteristics and we are **detecting object motion** when regions that exhibit motion characteristics are surrounded by regions that exhibit rest characteristics. Deformation characteristics may lend further support to the tracking hypothesis of such regions.

B.2 Motion Primitives

Depending on the outcome of this initial analysis the definitions for *onset, offset, occlusion*, and *disocclusion* are as follows:

- **Onset of a Motion**

In stationary mode: A region that changes from rest to motion characteristics.

In tracking mode: A region that changes from motion to rest characteristics.

- **Offset of a Motion**

In stationary mode: A region that changes from motion to rest characteristics.

In tracking mode: A region that changes from rest to motion characteristics.

- **Temporary Occlusion**

In stationary mode: An offset followed by a rest period followed by an onset.

In tracking mode: An onset followed by a motion period followed by an offset.

- **Temporary Disocclusion**

In stationary mode: An onset followed by a motion period followed by an offset.

In tracking mode: An offset followed by a rest period followed by an onset.

- **Triggering a Motion**

In stationary mode: Moving region 1 approaches resting region 2 and causes an onset.

In tracking mode: Moving region 1 approaches resting region 2 and causes an offset.

- **Ending a Motion**

In stationary mode: Moving region 1 approaches moving region 2 and causes an offset.

In tracking mode: Moving region 1 approaches resting region 2 and causes an onset.

References

[1] A. Akutsu and Y. Tonomura, "Video Tomography: An Efficient Method for Camerawork Extraction and Motion Analysis," in Proceedings of *Multimedia 1994*, ACM, 1994.

[2] F. Arman, R. Depommier, A. Hsu, and M.-Y. Chiu, "Content-based Browsing of Video Sequences," *Proc. ACM Multimedia*, pp. 97-103, 1994.

[3] V. Athitsos, M.J. Swain, and C. Frankel, "Distinguishing Photographs and Graphics on the World Wide Web," *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, pp. 10-17, 1997.

[4] N. Ayache and O.D. Faugeras, "Artificial Vision for Mobile Robots: Stereo Vision and Multisensory Perception,", MIT Press, 1991.

[5] R.A. Brooks, "A Robust Layered Control System for a Mobile Robot," A.I. Memo 864, MIT, 1985.

[6] J.F. Canny, "Finding Edges and lines in Images," Masters Thesis, MIT Press, 1983.

[7] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "A Fully Automated Content Based Video Search Engine Supporting Spatio-Temporal Queries," *IEEE Trans. Circuits and Systems for Video Technology*, 1998.

[8] B.B. Chaudhuri, N. Sarkar, and P. Kundu, "Improved Fractal Geometry Based Texture Segmentation Technique," *IEE Proceedings*, part E, vol. 140, pp. 233-241, 1993.

[9] R.W. Conners, C.A. Harlow, "A Theoretical Comparison of Texture Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 2, no 3, pp. 204-222, 1980.

[10] J. D. Courtney, "Automatic Video Indexing via Object Motion Analysis," *Pattern Recognition*, vol. 30, no. 4, pp. 607-626, 1997.

[11] G. Cybenko, "Approximation by Superposition of Sigmoidal Function," *Mathematics of Control, Signals, and Systems*, Chapter 2, pp. 303-314, 1989.

[12] A. Del Bimbo, E. Vicario, D. Zingoni, "A Spatial Logic for Symbolic Description of Image Contents," *J. Visual Languages and Computing*, vol. 5, pp. 267-286, 1994.

[13] N. Dimitrova and F. Golshani, "Motion Recovery for Video Content Classification," *ACM Trans. Information Systems*, vol. 13, no 4, pp 408-439, 1995.

[14] P. England, R.B. Allen, M. Sullivan, and A. Heybey, "I/Browse: The Bellcore Video Library Toolkit," *SPIE Proc. Storage and Retrieval for Image and Video Databases*, pp. 254-264, 1996.

[15] S. Fahlman, "Faster-Learning Variations on Back-Propagation: An Empirical Study," *Proc. Connectionist Models Summer School*, Morgan Kaufmann, 1988.

[16] O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, MIT Press, 1993.

[17] I.Fogel and D.Sagi, "Gabor Filters as Texture Discriminator," *J. Biological Cybernetics*, vol. 61, pp. 103-113, 1989.

[18] W.T. Freeman and E.H. Adelson, "The Design and Use of Steerable Filters," *PAMI*, Vol. 13, pp. 891-906, 1991.

[19] D. Gabor, "Theory of communication," *J. IEE*, vol. 93, pp. 429-457, 1946.

[20] N. Haering, Z. Myles, and N. da Vitoria Lobo, "Locating Deciduous Trees," *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, pp. 18-25, 1997.

[21] R.M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. Systems Man and Cybernetics*, vol. 3, no 6, pp. 610-621, 1973.

[22] M.R. Naphade, T. Kristjansson, B. Frey, and T.S. Huang, "Probabilistic Multimedia Objects (Multijects): A Novel Approach to Video Indexing and Retrieval in Multimedia Systmes", *Proceedings of ICIP '98*, 1998.

[23] S. S. Intille, "Tracking Using a Local Closed-World Assumption: Tracking in the Football Domain," *Master Thesis, M.I.T. Media Lab*, 1994.

[24] G. Iyengar and A. Lippman, "Models for Automatic Classification of Video Sequences", *SPIE Proc. Storage and Retrieval for Image and Video Databases*, pp. 216-227, 1997.

[25] R. Jain, R. Kasturi and B. Schunck, "Machine Vision," *McGraw Hill*, pp. 278-284, 1995.

[26] T. Kawashima, K. Tateyama, T. Iijima, and Y. Aoki, "Indexing of Baseball Telcast for Content-based Video Retrieval," *Proc. International Conference on Image Processing*, pp. 871-875, 1998.

[27] J.M. Keller and S. Chen, "Texture Description and Segmentation through Fractal Geometry," *Computer Vision, Graphics and Image Processing*, vol. 45, pp. 150-166, 1989.

[28] R. L. Lagendijk, A. Hanjalic, M. Ceccarelli, M. Soletic, and E. Persoon, "Visual Search in a SMASH System", *Proc. International Conference on Image Processing*, pp. 671-674, 1997.

[29] B.Manjunath and W.Ma, "Texture Features for Browsing and Retrieval of Image Data," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837-859, 1996.

[30] J. R. Miller, J. R. Freemantle, M. J. Belanger, C. D. Elvidge and M. G. Boyer, "Potential for determination of leaf chlorophyll content using AVIRIS", *Proceedings of the Second Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Workshop*, pp. 72-77, June 4-8, 1990, Pasadena, Calif. USA.

[31] Y. Awaya, J. R. Miller and J. R. Freemantle, "Background Effects on Reflectance and Derivatives in an Open-

Canopy Forest using Airborne Imaging Spectrometer Data", *Proceedings of the XVII Congress of ISPRS* Aug. 2-14, 1992 Washington, D.C., USA. pp. 836-843.

[32] M. Hebert, J. Ponce, T.E. Boult, A. Gross, and D. Forsyth, "Report of the NSF/ARPA Workshop on 3D Object Representation for Computer Vision", Dec. 5-7, 1994.

[33] S. Peleg, J. Naor, R. Hartley, and D. Avnir, "Multiple Resolution Texture Analysis and Classification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, no 4, pp. 518-523, 1984.

[34] A.P. Pentland, "Fractal-based Description of Natural Scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, no 6, pp. 661-674, 1984.

[35] P. Perona, "Deformable Kernels for Early Vision," *PAMI*, Vol. 17, pp. 488-499, 1995.

[36] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker, "Query by Image and Video Content: The QBIC System," *IEEE Computer*, Vol. 28, No. 9, pp. 23-32, September, 1995.

[37] R. J. Qian, M. I. Sezan and K. E. Matthews, "A Robust Real-Time Face Tracking Algorithm", *Proc. International Conference on Image Processing*, pp. 131-135, 1998.

[38] D. Saur, Y.-P. Tan, S.R. Kularni, and P.J. Ramadge, "Automated Analysis and Annotation of Basketball Video," *SPIE Proc. Storage and Retrieval for Image and Video Databases*, pp. 176-187, 1997.

[39] M. Smith and T. Kanade, "Video Skimming for Quick Browsing Based on Audio and Image Characterization," *CMU Computer Science Department Technical Report CMU CS-95-186*, 1995.

[40] M. Szummer, "Temporal Texture Modeling," *Master Thesis, M.I.T. Media Lab*, 1995.

[41] M. Szummer and R.W. Picard, "Indoor-outdoor image classification," in *IEEE Workshop on Content based Access of Image and Video Databases*, in conjunction with ICCV'98, (Bombay, India), Jan. 1998. http://www-white.media.mit.edu/people/szummer/profile.html

[42] A. Gupta and R. Jain, "Visual information retrieval", *Comm. Assoc. Comp. Mach.*, 40(5), May 1997

[43] A. Vailaya, A. Jain, and H.J. Zhang, "On Image Classification: City Images vs. Landscapes," *Workshop on Content based Access of Image and Video Libraries*, June, 1998.

[44] N. Vasconcelos and A. Lippman, "A Bayesian Framework for Semantic Content Characterization," *Proc. Computer Vision and Pattern Recognition*, pp. 566-571, 1998.

[45] J.S. Weszka, C.R. Dyer, and A. Rosenfeld, "A Comparative Study of Texture measures for Terrain Classification," *IEEE Trans. Systems Man and Cybernetics*, vol. 6, no 4, pp. 269-285, 1976.

[46] R.R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*, Statistical Modeling and Decision Science Series, Academic Press, 1997.

[47] M. Yeung, and B.-L. Yeo, "Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content," *IEEE Trans. Cicuits and Systems for Video Technology*, vol. 7, no 5, pp. 771-785, 1996.

[48] D. Yow, B.L.Yeo, M. Yeung, and G. Liu, "Analysis and Presentation of Soccer Highlights from Digital Video," *Proc. Asian Conference on Computer Vision*, 1995.

[49] H. J. Zhang, S. W. Smoliar, and J. H. Wu, "Content-Based Video Browsing Tools," *SPIE Proc. Storage and Retrieval for Image and Video Databases*, pp. 389-398, 1995.