A Vision System for Horizon Tracking and Object Recognition for Micro Air Vehicles

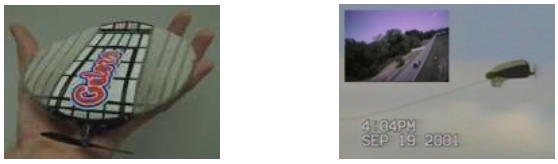
Sinisa Todorovic and Michael C. Nechyba
Department of Electrical and Computer Engineering
University of Florida
Gainesville, Florida 32611-6200
Email: {sinisha, nechyba}@mil.ufl.edu

***Abstract*— In this paper, we develop a unified vision system for small-scale aircraft, known broadly as Micro Air Vehicles (MAVs), that not only addresses basic flight stability and control, but also enables more intelligent missions, such as ground object recognition and moving-object tracking. The proposed system defines a framework for real-time image feature extraction, horizon detection and sky/ground segmentation, and contextual ground object detection. Multiscale Linear Discriminant Analysis (MLDA) defines the first stage of the vision system, and generates a multiscale description of images, incorporating both color and texture through a dynamic representation of image details. This representation is ideally suited for horizon detection and sky/ground segmentation of images, which we accomplish through the probabilistic representation of tree-structured belief networks (TSBN). Specifically, we propose incomplete meta TSBNs (IMTSBN) to accommodate the properties of our MLDA representation and to enhance the descriptive component of these statistical models. In the last stage of the vision processing, we seamlessly extend this probabilistic framework to perform computationally efficient detection and recognition of objects in the segmented ground region, through the idea of visual contexts. By exploiting the concept of visual contexts, we can quickly focus on candidate regions, where objects of interest may be found, and then compute additional features through the Complex Wavelet Transform (CWT) and HSI color space for those regions, only. These additional features, while not necessary for global regions, are useful in accurate detection and recognition of smaller objects. Throughout, our approach is heavily influenced by real-time constraints and robustness to transient video noise.**

I. Introduction

Over the past several years, researchers at the University of Florida have established an active program in developing *Micro Air Vehicles* (MAVs); for example, Figure 1 shows one of our current MAVs. More recently, we have demonstrated vision-based flight stabilization of MAVs [1], [2]. This paper moves well beyond this previous work in considering robust stabilization for less benign environments, and in developing a unified vision-based framework for control, tracking and recognition of ground objects to enable intelligent mission profiles.

Imparting MAVs with autonomous and/or intelligent capabilities is a difficult problem, due to their stringent payload requirements and the relative inadequacy of currently available miniature sensors suitable for these small-scale flight vehicles. The one sensor that is absolutely critical for almost any potential MAV mission (e.g. remote surveillance), however, is an on-board video camera. While an on-board camera is essential for a remote human supervisor, it also offers a wealth of information that can be exploited in vision-based algorithms for guidance and control of MAVs. As such, this paper presents a unified computer vision framework for MAVs that not only addresses basic flight stability and control, but also enables more intelligent missions, such as ground object recognition and moving-object tracking. In this framework, we first seek to extract relevant features from the flight video that will enable higher level goals. Then, we apply this image representation towards horizon detection and sky/ground segmentation for basic flight stability and control. Finally, we extend this basic framework through the idea of visual contexts to perform object recognition in the flight video. Throughout, the most important factors that inform our design choices for the vision system are: (1) real-time constraints, (2) robustness to video noise, and (3) complexity of various object appearances in flight images.



(a) (b)

Fig. 1. (a) six-inch UF MAV, (b) six-inch MAV in flight.

Subject to lighting conditions, landscape and video noise, recognition of various appearances in natural-scene images is a challenging problem even for the human eye. Therefore, we resort to a probabilistic formulation of the problem, where careful consideration must be given to selecting sufficiently discriminating features and a sufficiently expressive modeling framework. In our approach, feature extraction is performed by multiscale linear-discriminant analysis (MLDA), which provides a harmonic, multiscale representation of images [3]–[5]. MLDA explicitly incorporates color in its feature representation, while implicitly encoding texture through a dynamic representation of image details. Not only have we shown that MLDA-based horizon tracking runs in real-time, but MLDA's sparse representation of images also significantly speeds up computation in subsequent stages of the overall vision system.

Having detected the horizon, we proceed with sky/ground supervised segmentation, resorting to a Bayesian formulation,

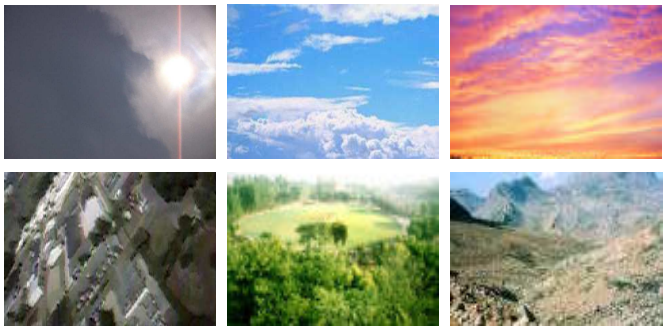


Fig. 2. Sample training images for the sky and ground classes.

where image classes are represented by statistical models. Although modeling sky and ground may seem intuitively easy, it is, in fact, a very challenging task. To statistically model complex variations in these two image classes, as illustrated in Figure 2, we propose the novel incomplete meta tree-structured belief network (IMTSBN). IMTSBNs are capable of representing neighborhoods in the image at varying scales, mirroring the MLDA multiscale image representation. Moreover, IMTSBNs have efficient linear-time inference algorithms.

Sky/ground segmentation offers important *contextual* priors to further segmentation and detection algorithms. In our approach to object recognition, we exploit the idea of *visual contexts* [6], where initial identification of the overall type of scene facilitates recognition of specific objects/structures within the scene. Thus, objects (e.g., cars, buildings), the locations where objects are detected (e.g., road, meadow), and the category of locations (e.g., sky, ground) form a taxonomic hierarchy. To account for different flight scenarios, different sets of image classes should be defined accordingly. Here again, we formulate object recognition as a supervised learning problem, building IMTSBNs for an extended set of image classes. However, while MLDA is sufficient for reliable recognition of global regions, it yields poor results for smaller objects. Therefore, in the last stage of the object recognition algorithm, once candidate object locations are detected using MLDA, we resort to an additional feature set–namely, the Complex Wavelet Transform (CWT) [7] and HSI color values. CWT and HSI coefficients form complete quad-tree structures for which we then build tree-structured belief networks (TSBNs) and ultimately perform supervised pixel labeling.

Our paper is organized as follows. In Section II, we define MLDA and motivate its applicability to our system. Next, in Section III, we show how MLDA naturally allows us to perform horizon detection. Then, in Section IV, present the IMTSBN model and discuss its appropriateness for the problem of sky/ground segmentation. The context-based object recognition algorithm is explained in Section V. Finally, in Section VI, we demonstrate the performance of the proposed unified computer vision system.

II. Multiscale linear-discriminant analysis

Our choice of feature selection is largely guided by extensive experimentation [8], [9], from which we conclude that a prospective feature space must span both color and texture domains. We have considered numerous image analysis tools discussed in the image processing literature. For instance, in [10], wavelet-domain analysis of natural-scene images was shown to be very efficient, especially in the presence of video noise. On the other hand, recent findings on human vision and natural image statistics [11], [12] undermine the appropriateness of wavelet-based approaches. In this literature, it has been reported that unlike wavelets, cortical cells are highly sensitive to orientation and elongation, along with the location and scale of stimuli. Moreover, the basis elements which best "sparsify" natural scenes are, unlike wavelets, highly direction-specific. Hence, recently many new image analysis methods, such as wedgelets, ridgelets, beamlets, etc. [4], [13] have been proposed. Aside from the multiscale and localization properties of wavelets, these methods exhibit characteristics that account for concepts beyond the wavelet framework, as does our approach to feature extraction and image representation–namely, *Multiscale Linear Discriminant Analysis (MLDA)* [3]. Not only does MLDA overcome the shortcomings of wavelets, but it also incorporates color information (unlike wavelets), satisfying the aforementioned requirements for the feature space.

An MLDA atom w is a piecewise constant function on either side of a linear discriminant that intersects a square in vertices v_i and v_j, as illustrated in Figure 3a. The discriminant (v_i, v_j) divides the square into two regions characterized by μ_0 and μ_1, which represent the mean vectors of RGB pixel color values for the two regions. Decomposing the image into a number of dyadic squares and finding their corresponding MLDA atoms, we obtain the MLDA dictionary over a finite range of locations, orientations and scales, as shown in Figure 3b.

MLDA image-feature extraction is performed by searching through the MLDA dictionary for the atoms that best represent the analyzed image with respect to two competing criteria: *discrimination* and *parsimony*. With respect to discrimination, we seek the direction (v_i, v_j), characterized by the maximum Mahalanobis distance J between two regions in a square, where

$$(\hat{v}_i, \hat{v}_j): \; J = \max_{(v_i, v_j)} \{(\mu_0 - \mu_1)^T (\Sigma_0 + \Sigma_1)^{-1} (\mu_0 - \mu_1)\}, \quad (1)$$

and Σ_0 and Σ_1 denote RGB covariance matrices of the two regions. Note that the computational cost of an exhaustive

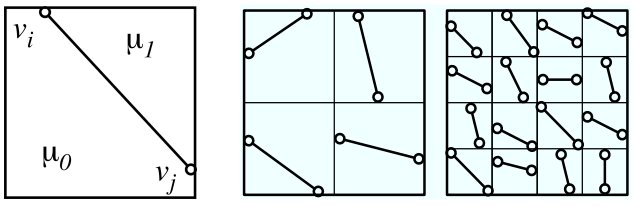


Fig. 3. (a) MLDA atom; (b) MLDA dyadic decomposition.

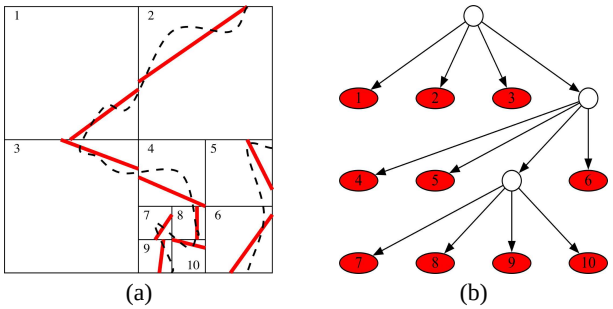


Fig. 4. (a) MLDA graph: the dashed line depicts the actual curve; (b) corresponding MLDA tree: ellipsoid nodes represent the leaf MLDA atoms.

search over a finite set of linear discriminants $\{(v_i, v_j)\}$ can be reduced by updating the relevant statistics only with the pixel values of delta regions (areas between two consecutive candidate linear discriminants). As the size of squares decreases, we achieve better piece-wise linear approximation of boundaries between regions in the image, as illustrated in Figure 4. Thus, the analyzed image is decomposed into dyadic squares organized in the MLDA tree $\mathcal{T}$. The MLDA tree $\mathcal{T}$ is *incomplete*, because atom generation stops at different scales for different locations in the image. The leaf nodes of $\mathcal{T}$ store the final MLDA image representation.

To control the generation of children dyadic squares, we impose the second optimization criterion (i.e. parsimony), as a counter-balance to accuracy. We define the cost function $R(\mathcal{T})$ to measure the parsimony of $\mathcal{T}$,

$$R(\mathcal{T}) = \sum r(w_\ell) + \alpha|\mathcal{T}| \quad , \tag{2}$$

where $r(w_\ell)$ is the inverse of the Mahalanobis distance computed for the corresponding leaf node w_ℓ, $r(w_\ell) = 1/J$, $|\mathcal{T}|$ denotes the number of terminal nodes in $\mathcal{T}$, and α represents the complexity cost per terminal node. Clearly, an exhaustive search in tree space for the minimum cost function is computationally prohibitive. Therefore, we implement a one-step optimization procedure, as in [14].

Instead of stopping at different terminal nodes, we continue the MLDA decomposition until all leaf squares are small in size, resulting in a large tree. Then, we selectively prune this large tree upward using the cost function $R(\mathcal{T})$. From expression (2), it follows that we can regulate the pruning process by increasing α to obtain a finite sequence of subtrees with progressively fewer leaf nodes. First, for each node $w \in \mathcal{T}$, we determine α_w for which the cost of a subtree $\mathcal{T}_w$ is higher than the cost of its root node w, as follows:

$$\begin{aligned} R(\mathcal{T}_w) \geq R(w) \;&\Rightarrow\; \sum r(w_\ell) + \alpha_w|\mathcal{T}_w| \geq r(w) + \alpha_w \cdot 1\,, \\ &\Rightarrow\; \alpha_w = \frac{r(w) - \sum r(w_\ell)}{|\mathcal{T}_w| - 1}\,. \end{aligned} \tag{3}$$

Then, the whole subtree $\mathcal{T}_w$ under the node w with the minimum value of α_w is cut off. This process is repeated until the actual number of leaf nodes is equal to or less than the desired number of leaf nodes. We note that the pruning process described above is computationally fast and requires only a small fraction of the total tree construction time. Starting with a complete tree, the algorithm initially trims off large subtrees with many leaf nodes. As the tree becomes smaller, the procedure tends to cut off fewer nodes at a time.

MLDA implicitly encodes information about spacial frequencies in the image (i.e. texture) through the process of tree pruning. Furthermore, the MLDA tree can easily be examined for spacial interrelationships of its linear discriminants, such as connectedness, collinearity and other properties of curves in images. Therefore, the MLDA representation is appropriate for computer-vision tasks where both color and texture are critical features.

III. Horizon-detection algorithm

The MLDA framework presents an efficient solution to the horizon-detection problem, incorporating our basic assumptions that the horizon can be interpolated by piecewise linear approximations (i.e., linear discriminants), and that the horizon separates the image into two homogeneous regions. First, the MLDA image representation is found for the cost function R, as defined by (2). The precise value of R is largely dependent on the video transmission quality for a given location, weather conditions, etc. Then, linear discriminants are extracted by exhaustively analyzing all possible directions, determined by a finite set of points $\mathcal{V}$ along the perimeter of the image. Each pair of points $(v_i, v_j) \in \mathcal{V}$ defines a direction with a slope a_{ij} along which MLDA atoms are examined. If a_{ij} coincides with the slope of an MLDA discriminant (within a predefined margin), then that linear discriminant is extracted as a part of the analyzed direction (v_i, v_j). Finally, among the candidate sets of the extracted MLDA discriminants, we choose the direction $(\hat{v}_i, \hat{v}_j)$ for which the sum of Mahalanobis distances J, as computed in (1), is maximum.

In benign flight video, the horizon-detection problem can be adequately solved by computing only the root atom of the MLDA tree, where the linear discriminant of the root is the optimal solution for the horizon estimate, as illustrated in Figure 5; in fact, the work presented in [1], [2] is equivalent to this special case. There are, however, unfavorable image conditions (e.g., when the horizon is not a straight line and/or an image is corrupted by video noise), when it is necessary to examine discriminants at finer MLDA scales. For example, in Figure 6, we show that the discriminant of the root MLDA atom does not coincide with the true horizon due to video noise. Expanding the MLDA tree corresponds to image filtering and leads to more accurate positions for the linear discriminants, which

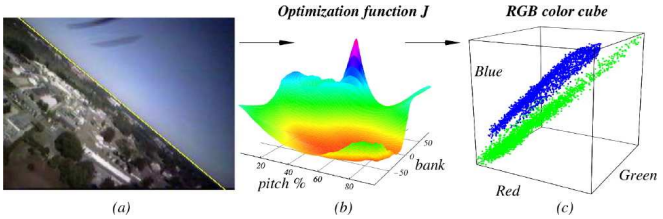


Fig. 5. (a) original image; (b) optimization criterion J as a function of bank angle and pitch percentage, that is, vertices along the perimeter of the image; (c) resulting classification of sky and ground pixels in RGB space.

(a) (b) (c)

Fig. 6. MLDA efficiently filters video noise: (a) noise degraded original image; (b) MLDA root atom with the discriminant not equal to the true horizon; (c) MLDA atoms at finer scales as clues for the true horizon position.

(a) (b) (c)

Fig. 7. Non-straight-line horizon: (a) original image; (b) the root MLDA atom with the discriminant not equal to the true horizon; (c) MLDA atoms at finer scales as clues for the true horizon position.

then give more accurate evidence of the true horizon's location in the image. Also, in Figure 7, we show an example, where MLDA detects the true horizon more correctly, as compared to a single-line discriminant. Additional examples and videos can be found at `http://mil.ufl.edu/~nechyba/mav`.

Finally, to accomplish horizon tracking in a sequence of flight images, we conduct a full search with every frame, rather than just a partial search in the neighborhood of the previous frame's horizon estimate. An error in the horizon estimate at time t_0 may permanently lock us into incorrect horizon estimation, with potentially catastrophic results, as discussed in greater detail in [2].

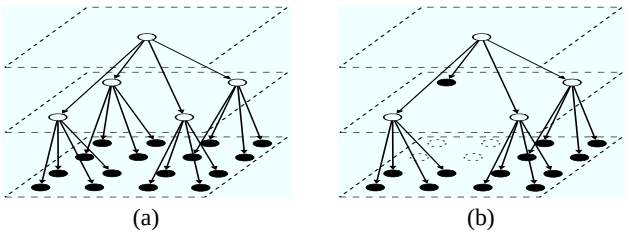
IV. Sky/Ground Segmentation

The horizon detection algorithm determines only the line separating sky and ground regions, which proves sufficient for vision-based flight control under benign flight conditions [1], [2]. However, for complex mission scenarios, correct identification of the sky and ground regions, rather than just the line separating them, takes on increased importance. To accomplish reliable sky/ground image segmentation, we resort to a Bayesian framework, choosing tree-structured belief networks (TSBNs) as the underlying statistical models to describe the sky and ground classes. There are several reasons for this choice of model. Successful sky/ground segmentation implies both accurate classification of large regions, as well as distinct detection of the corresponding boundaries between them. To jointly achieve these competing goals, both large and small scale neighborhoods should be analyzed; this can be achieved through the multiscale structure of TSBNs that naturally arises from our multiscale image representation (i.e., MLDA). Further, it is necessary to employ a powerful statistical model that is able to account for enormous variations in sky and ground appearances, due to video noise, lighting, weather, and landscape variability. Prior work presented in the computer vision literature clearly suggests that TSBNs possess sufficient expressiveness for our goals. For instance, in [15] TSBNs are employed for segmentation of man-made structures in natural scene images and in [16] TSBNs are used for multiscale document segmentation. Occasionally, TSBNs give rise to blocky segmentations, which reflect the fixed-structure deficiency of TSBNs. Nevertheless, considering our stringent real-time constraints, the efficiency of TSBN inference algorithms outweighs this drawback.

A tree-structured belief network (TSBN) is a generative model comprising hidden and observable random variables (RVs) organized in a tree structure $\mathcal{T}$. In supervised learning problems, such as our formulation of the sky/ground segmentation problem, an observable RV y represents an extracted feature vector (i.e., MLDA atom in our case) and the corresponding hidden RV x identifies the image class, $k \in C$ of y. As is common for TSBNs, the edges between the nodes, representing x's, describe Markovian dependencies across the scales, whereas y's are assumed mutually independent given the corresponding x's, as depicted in Figure 8.

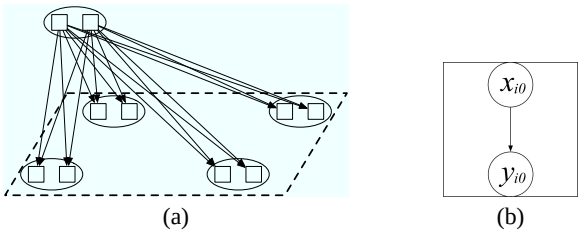
There are, however, substantial differences between our approach and the standard TSBN treatment in the graphical-models literature (e.g., [15], [17]). First, our TSBN model corresponds one-to-one to the MLDA tree. Therefore, due to MLDA-tree pruning, leaf nodes of a TSBN may occur at coarser resolutions as well as at the pixel level, resulting in an incomplete tree structure (see Fig. 8b). Second, we enhance the descriptive component of TSBNs by incorporating observable information at all levels of the TSBN model, not only at the lowest pixel level. Thus, to each MLDA atom w_i in the MLDA tree, we assign a pair of observable RVs, (y_{i0}, y_{i1}), modeling μ_0 and μ_1 RGB pixel color values. Clearly, the corresponding pair of hidden RVs, (x_{i0}, x_{i1}), models the image classes of μ_0 and μ_1. Third, Markovian dependencies among x's across scales model the relation that a parent MLDA atom generates four children MLDA atoms. Given that i is a child of j, it follows that both x_{i0} and x_{i1} have the same parents x_{j0} and x_{j1}, as illustrated in Figure 9a. Therefore, our statistical model is not a tree in the strict sense of the word; however, pairs (x_{i0}, x_{i1}), $\forall i \in \mathcal{T}$, form a tree structure, allowing us to use the belief-network formalism. To emphasize all the aforementioned differences between our model and "standard" TSBNs, we refer to our model as incomplete meta tree-structured belief networks (IMTSBN).

The state of a node, x_{ia}, is conditioned on the state of its



(a) (b)

Fig. 8. Differences in TSBN models: (a) "standard" complete TSBN; (b) incomplete TSBN: leaf nodes (depicted black) may occur at coarser resolutions due to MLDA-tree pruning.



(a) (b)

Fig. 9. (a) Zoomed-in look at Fig. 8b: statistical dependencies among parent and children nodes; (b) zoomed-in square from (a): observables are mutually independent given the corresponding hidden RVs.

parents, x_{jb}, and is given by conditional probability tables, P_{ijab}^{kl}, $\forall i,j \in \mathcal{T}$, $a,b \in \{0,1\}$, $\forall k,l \in C$. It follows that the joint probability of all hidden RVs, $X = \{x_{ia}\}$, can be expressed as

$$P(X) = \prod_{i,j \in \mathcal{T}} \prod_{a,b \in \{0,1\}} \prod_{k,l \in C} P_{ijab}^{kl} \, . \tag{4}$$

We assume that the distribution of an observable RV, y_{ia}, depends solely on the node state x_{ia}. Consequently, the joint pdf of $Y = \{y_{ia}\}$ is expressed as

$$P(Y|X) = \prod_{i \in \mathcal{T}} \prod_{a \in \{0,1\}} \prod_{k \in C} p(y_{ia}|x_{ia} = k, \theta_i^k) \, , \tag{5}$$

where $p(y_{ia}|x_{ia} = k, \theta_i^k)$ is modeled as a mixture of M Gaussians whose parameters are grouped in θ_i^k. We usually simplify the notation as $p(y_{ia}|x_{ia} = k, \theta_i^k) = p(y_{ia}|x_{ia})$. For large M, the Gaussian-mixture density can approximate any probability density [18], in particular, the distributions of image classes. In order to avoid the risk of overfitting the models, we assume that $\theta's$ are equal for all i at the same scale. Finally, an IMTSBN is fully specified by the joint distribution of X and Y given by

$$P(X,Y) = \prod_{i,j \in \mathcal{T}} \prod_{a,b \in \{0,1\}} \prod_{k,l \in C} p(y_{ia}|x_{ia}) P_{ijab}^{kl} \, . \tag{6}$$

To learn the parameters of an IMTSBN, we iteratively maximize the conditional probability $P(X|Y)$, using a probabilistic inference algorithm for TSBNs, broadly known as Pearl's $\lambda - \pi$ message passing scheme [19], [20]. In the image-processing literature similar algorithms have been proposed [16], [21]. Essentially, all these algorithms perform belief propagation up and down the tree, where after a number of training cycles we obtain all the tree parameters necessary to compute $P(X|Y)$. Note that simultaneously with Pearl's belief propagation, we employ the EM algorithm [22] to learn the parameters of the Gaussian mixture distributions. Since IMTSBNs have observable RVs at all tree levels, the EM algorithm is performed at all scales. Moreover, Pearl's message passing scheme is accommodated to sum λ messages over eight children nodes for each parent when propagating upwards, and to account for π messages of two parents for each child node when propagating downwards (see Fig. 9a). Thus, the parameters of the prior models are learned through extensive training, which can be performed off-line[1]. Once learned, the parameters fully characterize the likelihoods of image classes, given the pixel values. These likelihoods are then submitted to a Bayes classifier to perform image segmentation.

[1]In this case, there are no real-time constraints for training.

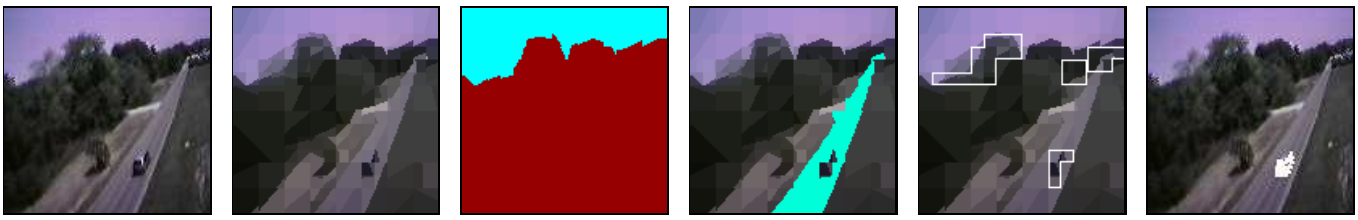
V. Object Recognition Using Visual Contexts

In our approach to object recognition, we seek to exploit the idea of *visual contexts* [6], and, thus, to incorporate the algorithms for flight stability and control (i.e., the MLDA-based horizon detection and sky/ground segmentation) into a unified framework. Having previously identified the overall type of scene, we can then proceed to recognize specific objects/structures within the scene. Thus, objects (e.g., cars, buildings), the locations where objects are detected (e.g., road, meadow), and the category of locations (e.g., sky, ground) form a taxonomic hierarchy. There are several advantages to this type of approach. Contextual information helps disambiguate the identity of objects despite the poverty of scene detail (flight images). Second, visual contexts significantly reduce the search required to locate specific objects of interest, obviating the need for an exhaustive search for objects over various scales and locations in the image. Finally, as discussed in Section II, our low-dimensional image representation is very resilient to video noise, yielding better object-detection results.

For each image in a video sequence, we first compute its MLDA representation and perform categorization–that is, sky/ground image segmentation. Then, we proceed with localization–that is, recognition of global objects/structures (e.g., road, forest) in the ground[2] region. Following the Bayesian formulation, as explained in Section IV, to each ground MLDA atom, w_i, we assign a pair of observable RVs, (y_{i0}, y_{i1}), which are fully specified by a pair of hidden RVs, (x_{i0}, x_{i1}) representing a predefined set of global objects that may appear in flight images. To account for different flight scenarios, different sets can be defined accordingly. Using the prior knowledge of a MAV's whereabouts, we can reduce the number of image classes, and, hence, computational complexity as well as classification error. Clearly, our approach to global-object recognition is reminiscent of building IMTSBNs for the sky and ground classes. Here, we extend the set of classes C, generalizing the meaning of classes to any global-object appearance. Thus, the results from Section IV are readily applicable. To learn all the model parameters, we employ Pearl's belief propagation together with the EM algorithm. Note that, for labeling only ground w's, we build an IMTSBN of smaller dimensions than the image size and, thus, achieve significant computational savings.

Loss of image detail, when computing the mean values of MLDA atoms, can hinder subsequent object recognition. While MLDA proves sufficient for reliable identification of global objects, it yields poor results for the recognition of small objects. Therefore, in the next step of the algorithm, our intention is not to actually recognize and differentiate between particular objects, but rather to detect candidate locations

[2]Recognition of objects in the sky region is of lesser importance to us; however, if required, it can be easily incorporated into the algorithm.



(a) (b) (c) (d) (e) (f)

Fig. 10. The hierarchy of visual contexts conditions gradual image interpretation: (a) a 64×64 MAV-flight image; (b) MLDA image representation; (c) categorization: sky/ground segmentation; (d) localization: recognition of global objects; (e) detection of candidate object locations; (f) car recognition using CWT and HSI features.

where objects of interest might appear in the image. For this task, it is necessary to resort to a more descriptive image representation than merely μ_0 and μ_1 of MLDA atoms. Hence, we examine spacial interrelationships of linear discriminants among neighboring MLDA atoms, such as connectedness, collinearity, etc. In particular, if our goal is to detect artificial objects in the image then it seems reasonable to examine the geometric regularities of the extracted edges. Thus, scanning the identified ground region in the image, we analyze 3×3 boxes of leaf-level MLDA atoms, checking whether neighboring linear discriminants are parallel, perpendicular or collinear, though the size of the analyzed window and testing procedure could easily be tailored to other application requirements. The examined image areas with the spacial frequency of the aforementioned geometric properties higher than an application-dependent threshold are extracted as candidate locations for object appearances. Note that terminal MLDA atoms may occur at different scales of an MLDA tree, representing varying image areas, speeding up the overall search.

To further examine the candidate object locations from the previous step, it becomes necessary to consider image analysis tools other than MLDA. The results of our extensive experimentation [8], [9], where various feature extraction methods have been investigated, suggest that the Complex Wavelet Transform (CWT) [7] together with HSI color features yield robust and reliable object recognition. Thus, having detected candidate object locations, we proceed with new feature extraction of those particular image regions, only. CWT and HSI coefficients form complete quad-tree structures for which we then build TSBNs, similar to the procedure explained in Section IV. The off-line training of TSBNs is performed for a predefined set of image classes (over CWT and HSI features). For instance, for remote traffic monitoring, a global object *road* may induce a set of objects $\{$*car, cyclist, traffic sign*$\}$. Consequently, for object recognition, we, in fact, consider only a small finite number of image classes, which improves recognition results.

Following inference down the hierarchy of visual contexts, we gradually perform categorization, localization and object recognition, as demonstrated in Figure 10. For space reasons, we present only the recognition of a global object *road*, then, the detection of candidate object locations in the ground region, and final recognition of an artificial object *car*, though the algorithm performs simultaneous searches over a larger set of objects. Obviously, the proposed algorithm performs well despite video noise and the poverty of image detail in the flight image.

VI. Results

At different times of the day, under various weather conditions, and for different landscapes, we have gathered hours of MAV-flight video. These images, organized into training and test data-sets, have been used in several sets of experiments, which we report below.

First, the MLDA feature extraction and our horizon-detection algorithm has been demonstrated to run at 30Hz on an Athlon 2.4GHz PC with a down-sampled image resolution of 64×64. For the test images, representative samples of which are illustrated in Figure 5–7, our horizon-detection algorithm correctly identifies the horizon in over 99.9% of cases.

Second, for training IMTSBNs for the sky and ground image classes, we used two sets of 500 sky and ground training images. We carefully chose the training sets to account for the enormous variability within the two classes, as illustrated in Figure 2. For each image, first the MLDA representation was found and then the corresponding IMTSBN model was trained. The training time for 1000 images of both classes takes less than 40s on an 2.4GHz x86 processor. Correct segmentation takes only 0.03s–0.07s, depending on the number of levels in the MLDA tree, for 64×64 image resolution. Clearly, the algorithm runs faster for a small number of MLDA terminal nodes, but at the cost of increased segmentation error. Note that while we are very close to meeting our 30Hz processing goal, it is not crucial for sky/ground segmentation, as long as this segmentation is updated sufficiently often. In between segmentation updates, horizon detection suffices for flight stability and control. Having trained our sky and ground statistical models, we tested the segmentation algorithm on 300 flight images. For accuracy, we separated the test images into three categories of 100 images each: (I) easy-to-classify sky/ground appearances (e.g. clear blue sky over dark-colored land), (II) challenging images due to landscape variability, and (III) noise-degraded images. In Figure 11, we illustrate classification performance on representative test images from each category. To measure misclassification, we marked the "correct" position of the horizon for each image by visual

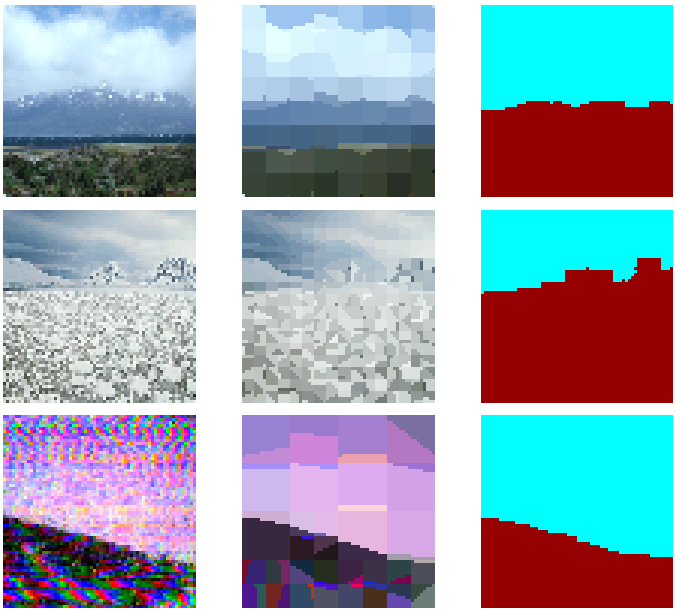


Fig. 11. Sky/ground segmentation of the three categories of test images: (top) mountain view (category I); (center) water surface covered with broken patches of ice similar in appearance to the sky (category II); (bottom) noise-degraded MAV flight image (category III).

inspection and then computed the percentage of erroneously classified pixels. In Table I, we summarize our segmentation results. Averaging results seemed inappropriate, because only a small number of test images in each category generated the larger error percentages.

Third, object recognition tests have been carried out for numerous types of objects, of which, for space reasons, we present only results for the following image classes: *road*, *car*, *cyclist*, *traffic sign*. For training TSBNs, we carefully chose 100 MAV-flight images for each image class. After experimenting with different image resolutions, we found that reliable object recognition was achievable for resolutions as coarse as 64×64 pixels. With this resolution, the processing time of the object-recognition procedure (i.e., sky/ground segmentation, road localization and car/cyclist/traffic-sign recognition), ranges from 1s to 4s on an Athlon 2.4GHz PC, depending on image complexity. Computing additional CWT and HSI features for candidate object locations is the most time consuming; therefore, we implement an optimization procedure that adaptively selects the optimal feature subset, as discussed in [9]. Thus, for simple images, where only one candidate location is detected, and where only one feature (e.g., wavelet coefficients oriented in the direction of $15°$) is used for pixel labeling, we achieve processing times of about 0.5s, which is sufficient for the purposes of moving-car or cyclist tracking. Moreover, for a sequence of images in video,

TABLE I

PERCENTAGE OF SKY/GROUND MISCLASSIFIED PIXELS

category I	category II	category III
2% - 6%	2% - 8%	5% - 14%

TABLE II

OBJECT RECOGNITION RESULTS

category I (20 objects)			category II (38 objects)			category III (20 objects)		
CR	SI	CM	CR	SI	CM	CR	SI	CM
16	2	2	31	3	4	11	2	7

TABLE III

PERCENTAGE OF CORRECTLY CLASSIFIED PIXELS FOR CR OUTCOMES

category I	category II	category III
94%	91%	87%

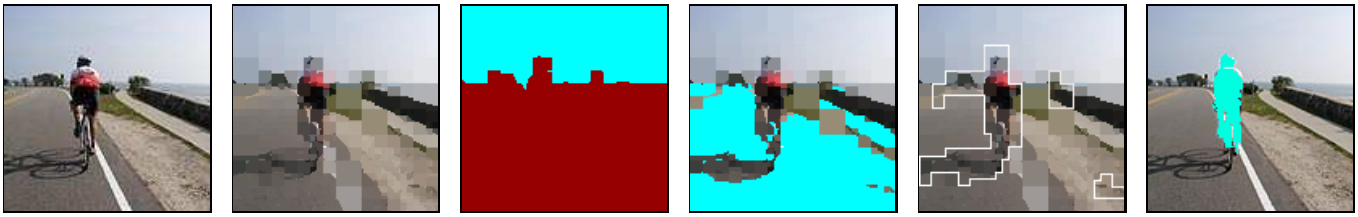
the categorization and localization steps could be performed only for images that occur at specified time intervals.

Similar to the sky/ground segmentation testing, we carried out object recognition experiments on three categories of test-sets of 20 images each, containing appearances of cars, cyclists and traffic signs, as illustrated in Figure 12 and Figure 13. By visual inspection, we hand-labeled the pixels belonging to objects for each test image. In Table II, we report the recognition results of 20, 38, and 20 objects in the first, second and third category of images, respectively. The outcomes are organized into three groups: correctly recognized objects (correct recognition – CR), detected object appearances with erroneous identification (swapped identity – SI), and undetected object appearances (complete miss – CM). For the CR group of outcomes, in Table III, we present the percentage of correctly classified pixels for the three categories of images. When considering the overall quality of recognition results, it is important to emphasize that we are dealing with relatively poor-quality images, not high-resolution, high-quality images. For example, the relatively larger error associated with category III images is obviously due to video-noise degradation.

VII. CONCLUSION

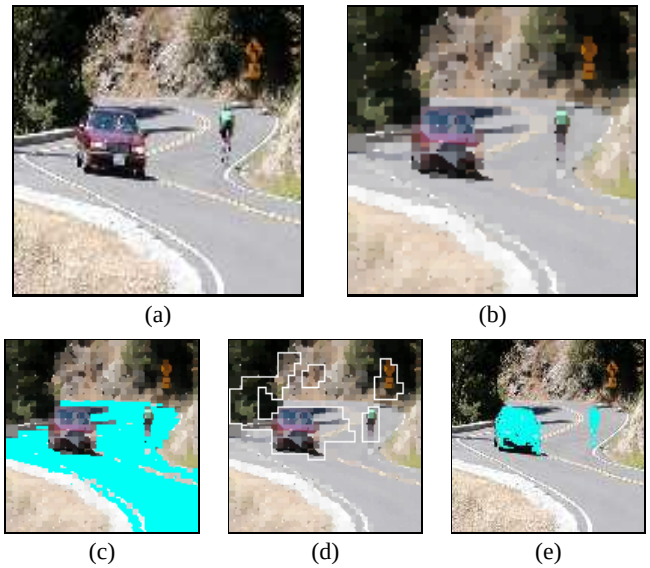
In this paper, we presented a unified computer-vision system for enabling intelligent MAV missions. We first developed MLDA as a feature extraction tool for subsequent image analysis. Next, we demonstrated MLDA's applicability to real-time horizon detection. Then, we explained the use of IMTSBN statistical models in segmenting flight images into sky and ground regions. We extended this basic framework through the idea of visual contexts to detect and recognize object/structure appearances. Finally, we presented the results of our extensive testing, which appear to validate our vision-system design choices. While testing to date has largely been done on recorded flight images and video, we are currently moving towards further closed-loop flight tests of the proposed contextual vision system.

The work presented in this paper is but one part of our on-going research effort to develop intelligent and mission-capable MAVs. Together with other researchers, we are also working on recovering 3D-scene structure of the ground from MAV flight videos. Additionally, work is on-going in sensor integration (e.g. GPS, INS) on our larger flight platforms (24" maximum dimension), flight vehicle modeling and characterization, and advanced control algorithm development. We



(a) (b) (c) (d) (e) (f)

Fig. 12. The hierarchy of visual contexts conditions gradual image interpretation: (a) a category I image; (b) MLDA image representation; (c) categorization: sky/ground segmentation; (d) localization: recognition of global objects; (e) detection of candidate object locations; (f) cyclist recognition using CWT and HSI features.



(a) (b) (c) (d) (e)

Fig. 13. Simultaneous recognition of objects: (a) a category II image; (b) MLDA image representation; (c) localization: recognition of global objects; (d) detection of candidate object locations; (e) car and cyclist recognition using CWT and HSI features. Note that while the traffic sign was detected as a candidate region, it was not recognized as such due to lack of detail.

expect that the totality of our efforts will eventually enable MAVs to not only fly in unobstructed environments, but also in complex 3D environments, such as urban settings.

REFERENCES

[1] S. M. Ettinger, M. C. Nechyba, P. G. Ifju, and M. Waszak, "Vision-guided flight stability and control for Micro Air Vehicles," in *Proc. IEEE Int'l Conf. Intelligent Robots and Systems (IROS)*, vol. 3, 2002, pp. 2134–40.

[2] ——, "Vision-guided flight stability and control for Micro Air Vehicles," *Advanced Robotics*, vol. 17, no. 7, pp. 617–40, 2003.

[3] S. Todorovic and M. C. Nechyba, "Multiresolution linear discriminant analysis: efficient extraction of geometrical structures in images," in *Proc. IEEE Int'l Conf. Image Processing*, vol. 1, 2003, pp. 1029–32.

[4] D.L.Donoho, "Wedgelets: Nearly-minimax estimation of edges," *Annals of Statistics*, vol. 27, no. 3, 1999.

[5] J.K.Romberg, M.Wakin, and R.G.Baraniuk, "Multiscale wedgelet image analysis: fast decompositions and modeling," in *Proc. IEEE Int'l Conf. Image Processing*, Rochester, NY, 2002.

[6] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," AI Laboratory – MIT, Tech. Rep., March 2003.

[7] N. Kingsbury, "Image processing with complex wavelets," *Phil. Trans. Royal Soc. London*, vol. 357, 1999.

[8] S. Todorovic, M. C. Nechyba, and P. G. Ifju, "Sky/ground modeling for autonomous MAVs," in *Proc. IEEE Int'l Conf. Robotics and Automation (ICRA)*, vol. 1, 2003, pp. 1422–7.

[9] S. Todorovic and M. C. Nechyba, "Towards intellignet mission profiles of Micro Air Vehicles: object recognition in flight images," to appear in *Proc. Eight European Conf. Comp. Vision*, May 2004.

[10] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain Hidden Markov Models," *IEEE Trans. Image Processing*, vol. 10, no. 7, 2001.

[11] S. issue, "Natural Stimulus Statistics," *Network: Computation in Neural Systems*, vol. 12, 2001.

[12] D.J.Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. of Optical Soc. of America, A*, vol. 4, 1987.

[13] A. G. Flesia, H. Hel-Or, E. J. C. A. Averbuch, R. R. Coifman, and D. L. Donoho, "Digital implementation of ridgelet packets," in *Beyond Wavelets*, J. Stoeckler and G. V. Welland, Eds. Academic Press, 2002.

[14] L.Breiman, J.Friedman, R.Olshen, and C.Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[15] S. Kumar and M. Hebert, "Man-made structure detection in natural images using a causal multiscale random field," in *Proc. IEEE Conf. Comp. Vision Pattern Rec.*, 2003.

[16] H. Cheng and C. A. Bouman, "Multiscale Bayesian segmentation using a trainable context model," *IEEE Trans. Image Processing*, vol. 10, no. 4, 2001.

[17] X. Feng, C. K. I. Williams, and S. N. Felderhof, "Combining belief networks and neural networks for scene segmentation," *IEEE Trans. Pattern Anal. Machine Intelligence*, vol. 24, 2002.

[18] M. Aitkin and D. B. Rubin, "Estimation and hypothesis testing in finite mixture models," *J. Royal Stat. Soc.*, vol. B-47, no. 1, 1985.

[19] J. Pearl, *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. San Mateo: Morgan Kaufamnn, 1988.

[20] B. J. Frey, *Graphical Models for Machine Learning and Digital Communication*. Cambridge, MA: The MIT Press, 1998.

[21] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using Hidden Markov Models," *IEEE Trans. Sig. Processing*, vol. 46, 1998.

[22] G. J. McLachlan and K. T. Thriyambakam, *The EM algorithm and extensions*. John Wiley & Sons, 1996.