# UAV Sensor Fusion with Latent-Dynamic Conditional Random Fields in Coronal Plane Estimation

Amir M. Rahimi*[1], Raphael Ruschel†[2] and B. S. Manjunath‡[1]

[1]Department of Electrical and Computer Engineering, University of California, Santa Barbara
[2]Universidade Estadual do Rio Grande do Sul, Porto Alegre, Brazil

## Abstract

*We present a real-time body orientation estimation in a micro-Unmanned Air Vehicle video stream. This work is part of a fully autonomous UAV system which can maneuver to face a single individual in challenging outdoor environments. Our body orientation estimation consists of the following steps: (a) obtaining a set of visual appearance models for each body orientation, where each model is tagged with a set of scene information (obtained from sensors); (b) exploiting the mutual information of on-board sensors using latent-dynamic conditional random fields (LDCRF); (c) Characterizing each visual appearance model with the most discriminative sensor information; (d) fast estimation of body orientation during the test flights given the LDCRF parameters and the corresponding sensor readings. The key aspects of our approach is to add sparsity to the sensor readings with latent variables followed by long range dependency analysis. Experimental results obtained over real-time video streams demonstrate a significant improvement in both speed (15-fps) and accuracy (72%) compared to the state of the art techniques that only rely on visual data. Video demonstration of our autonomous flights (both from ground view and aerial view) are included in the supplementary material.*

**Index Terms:** *micro-UAV, CRF, LDCRF, Coronal Plane, P-N learning, CRF Features, Cliques*

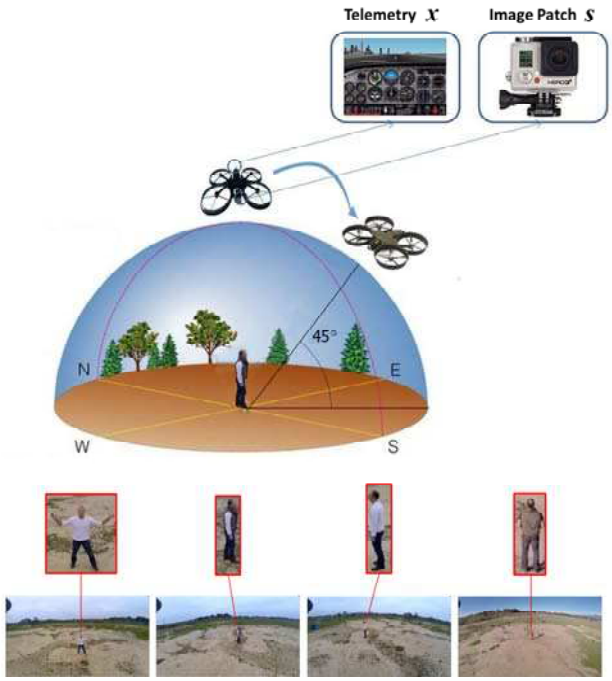


**Figure 1:** Overview of our UAV task. The visual appearance of each body orientation is characterized using latent-dynamic conditional random fields (LDCRF) over the scene characteristics. The scene characteristics are obtained from the UAV telemetry. Once the coronal plane is detected, the UAV automatically maneuvers to face the user.

## 1. Introduction

Robust human and robot interaction (HRI) methods are of essence in everyday activities. The performance of HRI systems relies heavily on accurate estimation of body orientation [3, 11]. A crucial requirement for active visual interactions with UAVs is to maneuver the UAV in front of a user. Facing the user is the first step towards effective gesture recognition algorithms [12]. However, techniques to detect body orientation from a mobile camera in natural environments are not applicable from UAV perspective. This is partially due to unexpected movements and particularly due to wide range of scene characteristics such as illumination, view angle and background differences. The general approach to addressing this problem involves extract-

*mohaymen@ece.ucsb.edu
†raphael-santos@uergs.edu.br
‡manj@ece.ucsb.edu

**Figure 2:** Examples of visual appearance for eight body orientation. Each row indicates the same body orientation for different persons, where the snapshots are taken at various telemetry settings.

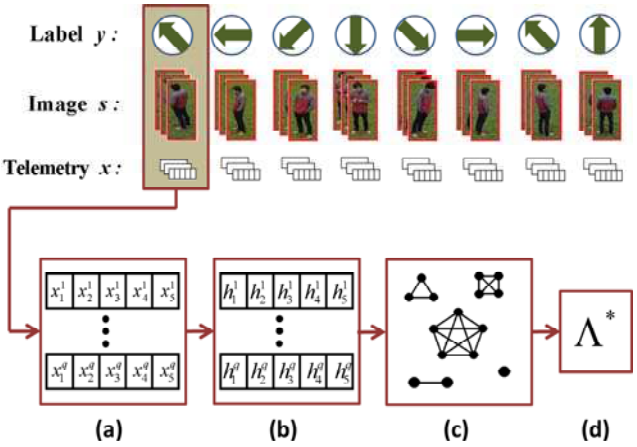


**Figure 3:** Training Process: The inputs is a set of image patches $\mathbf{s}$ from a given label (body orientation) $y \in \{y_1, ..., y_8\}$, where each patch is tagged with a set of telemetry data $\mathbf{x}$. The training process involves: a) Collecting a set of telemetry data for each label, b) Obtaining the latent state for each sensor reading, c) exploiting the most significant features using LDCRF and d) output a set of parameters $\Lambda^* = \{\lambda^1, ..., \lambda^5\}$, where each $\lambda^n$ is a set of optimal parameters from the $n^{th}$ order interaction.

ing the relevant low-level features/descriptors such as SIFT, HOG, Haar or local receptives, and applying classification schemes such as SVM or random forests or boosting methods [29, 8, 30, 4, 15, 10]. Some techniques articulate the "head" and "body orientation" to approximate the direction a person is facing. Others have leveraged prior information such as body shape to boost their performance [16, 14]. Researchers have also considered the relationship of head with respect to body in order to address this problem [7, 13, 5]. Given the nature of aerial videos (Figure 2); visual features alone are inadequate to distinguish between different appearance models (see results in section 4). However, as we demonstrate in the following, using visual appearance models together with scene context–as measured by the various non-visual sensors on-board the UAV, can significantly improve the overall efficiency and effectiveness. In our specific application, as illustrated in Figure 1, we infer the subject's coronal plane[1] in order to maintain the UAV position in front of the person.

In order to obtain a fast and robust visual appearance model of the user, many researchers have considered a set of positive and negative samples to retrain their model [23, 6, 17, 9]. Our approach is to use the *P-N learning* technique [20] where the models are template-based and in conjunction with the feedback mechanism introduced in [20], appearance models can be updated continuously during object tracking. However, a major shortcoming in the UAV application context is the requirement to evaluate a large number of templates over all possible body orientations for each scene characteristics. Instead, we propose to build a probabilistic framework using latent-dynamic conditional random fields (LDCRF) over the sensors to exploit the scene context and limit the search space. Specifically, the scene context include: Altitude (A), Gimbal Angle (G), UAV Magnetic Heading (H), Time of the Day (T), and GPS Location (L). The intuition is that visual appearance can vary significantly from one scene context to another, and this side information can facilitate the identification process. For example, when the Sun is behind a person, the silhouette of the person is significantly darker compared to a different viewing direction or if the same body orientation is observed from two different angle the visual appearance of the body looks very different (Figure 6). *Atomic* sources of information (A,G,H,T,L) and their combinations (i.e. *factors*) each contain a degree of information that influence the appearance models. From a graphical model perspective, the influence of each factor can be measured by the gain obtained in the likelihood models. In this paper, the mutual information between factors is measured in CRF framework. The CRF features are dynamically selected and weighted as new observations could potentially change the learned relationships. We regulate the computational complexity vs. likelihood gains using the Akaike Information Criterion [2]. During the test flight, the scene information is given by telemetry at no additional computational cost and is used to efficiently select the best model from a large set of potential candidates.

[1]The coronal plane is the physiological plane that splits a body into two parts: front and back and extends from top (head) to bottom (feet).

### 1.1. Contribution Summary

The main contributions of this work are:
**1.** Introducing a new technique that links the non-visual sensor information to the visual appearance of an object.
**2.** Exploiting the sensor information with LDCRF where high order interactions are evaluated using Akaike Information Criterion.
**3.** Present an autonomous UAV system capable of estimating the coronal plane based on scene characteristics and navigating the UAV to the frontal view.

## 2. Problem Statement

Let us consider an image patch $s^t$, which is obtained from a person-bounding box at $t^{th}$ instance. Each image patch $s^t$ is synchronized with a set of telemetry data $\mathbf{x}^t$ s.t. $\mathbf{x}^t = \{x_1^t, x_2^t, ..., x_M^t\}$. Given an instance $t$, our goal is to identify the label $y^t$ corresponding to the body orientation with respect to the UAV camera. This can be formulated as maximizing the correlation between the observation and a stored visual appearance template,

$$\hat{y} = \underset{t',y}{\operatorname{argmax}} \; n(s)^T n(\mu_{s^{t'}}, y) \tag{1}$$

where $\mu_{s^t}$ is a template based appearance model given the label $y$, and $n(.)$ is used to normalize the window size between the template and the image patch. The visual appearance models are generated every few frames. This leads to a large number of models to select from during the test flight. Since every model has a corresponding telemetry associated to it, the model selection can be projected over the telemetry data. Given the telemetry data observed at the scene, the goal then is to find an instance within the training set where the scene characteristics are most similar. Not all telemetry data components contribute equally to the model selection process and they are not mutually independent either given a label. In the following we propose an algorithm based on latent-dynamic conditional random fields (LDCRF) to model the implicit dependencies on the telemetry data towards efficient subset selection of visual appearance models (Figure 4).

## 3. Model Selection with LDCRF

Conditional random fields (CRFs) [22] are convenient models to encode dependencies in undirected graphical models. The power of CRF models rely upon the quantity and quality of feature functions. We briefly describe the template-based appearance model followed by a general formulation of CRF over the sensor data. The performance of CRF model is improved by adding latent variables $h_i$ to each local observation $x_i$, where the range of possible states are significantly reduced. Next, we dynamically select discriminative feature functions and hence the associated visual appearance models using the Akaike information criterion [2].

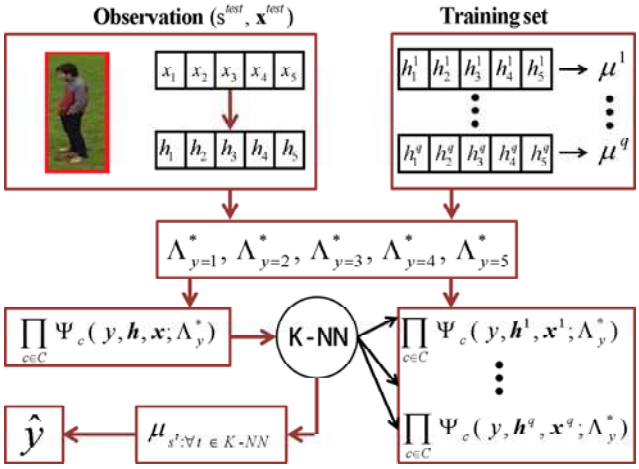


**Figure 4:** Testing Process: Given the weights learned during the training we compute the sum of potentials $\Psi(.)$ for every telemetry in the training set. During the test flight we compute the same sum of potentials for the given instance. $K$ nearest samples , from each orientation, are then selected and evaluated using equation (1). The model with the highest spatial correlation is then used to recognise the label (body orientation).

### 3.1. Visual Appearance Model with *P-N Learning*

We create a visual appearance model $\mu_{s^t}$ for each image patch $s^t$ and characterize each model with the corresponding $\mathbf{x^t}$. The visual appearance model $\mu_s$ consists of a few positive and negative patches obtained from the vicinity of the bounding box similar to [18]. The accurate selection of positive and negative patches are crucial to the quality of each model. Hence, each positive and negative samples are evaluated with *P-N learning* technique [19] to minimize the error. During test flight, given the observation $(\mathbf{s}, \mathbf{x})$, we first select the top candidate models for each label (the selection process is described in section 3.5) and among those that are selected, the appearance model that has the highest correlation with $\mathbf{s}$ determines the final label $\hat{y}$, as discussed in equation (1).

### 3.2. CRF Model to Exploit Telemetry

Consider a graphical model $G = (V, E)$, where $V$ is a set of $M$ nodes, each corresponding to a particular telemetry measure $x_i$. $E$ is a set of edges between different cliques. The cliques are defined with any combination of telemetry information $x_i$. Figure 5 illustrate the range of possible cliques in different colors. The relationship modelings in CRFs are typically limited to the second order interactions due to intractable inference techniques. However, in our case with only 5 measurements we are able to exploit all possible combinations with tractable inference. The UAV

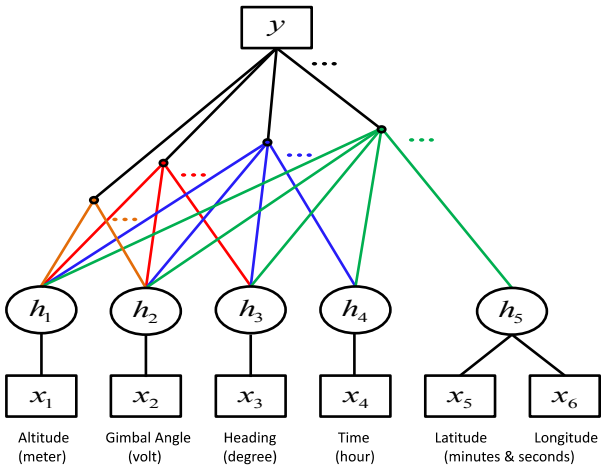


**Figure 5:** Structure of LCRF over five sensor information. The latent states are added to increase sparsity of local variables. All possible cliques of information is evaluated and the most discriminative ones are selected to improve speed and accuracy.

telemetry information consists of Altitude ($x_1$), Gimbal-Angle ($x_2$), Time-of-Day ($x_3$), Magnetic-Heading ($x_4$), and GPS-Location ($x_5$). Our objective here is to exploit the mutual information within $\{x_i\}$ to characterise each model given the labels (i.e. body orientation). This can be written as,

$$p(y|\mathbf{x};\theta) = \frac{1}{Z(\mathbf{x})} \prod_{c \in C} \Phi_c(y, \mathbf{x}; \theta) \tag{2}$$

where the clique potentials $\Phi_c(y, \mathbf{x})$ are non-negative real values obtained from the feature functions of each cliques and the normalizing term $Z(\mathbf{x})$, also referred as the partition function, is defined as $Z(\mathbf{x};\theta) = \sum_y \prod_{c \in C} \Phi_c(y, \mathbf{x}; \theta)$. Each clique potential is factorized over a set of feature functions $f_k^c(.)$, where $k$ is the index of feature functions in the clique $c$. The significance of each feature function is reflected in their parameters $\theta_k^c$. In the standard CRF terminology, the potentials are written as:

$$\Phi_c(y, \mathbf{x}; \theta) = \exp(\sum_k \theta_k^c f_k^c(y, \mathbf{x})) \tag{3}$$

The value of each parameter $\theta_k$ is directly influenced by the sparsity of the data (i.e. $\mathbf{x}$). Increasing the sparsity improves the expressive power of CRF structures, which is also the underlying rational in the case of Hidden Markov Models (HMM) or Dynamic Bayesian Networks (DBN). Literature has repeatedly shown that adding a set of latent variables can improve performance due to added sparsity (e.g. [26, 25]). In the next section, we also expand the formulation of CRF by assigning a single latent variable $h_i$ to each node $x_i$ such that $h_i = \Omega(x_i)$, where $\Omega(.)$ is an arbitrary mapping function used to add sparsity.

### 3.3. Increasing Sparsity with Latent CRF

One of the main advantages of CRF models is the flexibility of feature functions. Any possible pattern over the latent states can be considered. However, if the observed information is densely distributed, the feature functions tend to impose lesser discriminative value. In the case of telemetry information the scope of each local observation is too large. For instance the latitudes and longitudes obtained from the GPS node (i.e. $x_5$) is dense at less than a meter resolution. Formulating a dependency model over such a dense distribution is clearly ineffective. Instead, we introduce latent variables $h_i$ to each node $x_i$ such that $h_i = \Omega(x_i)$. The mapping function $\Omega(.)$ is arbitrary (e.g. clustering, rule-based techniques). In our setup we cluster each local observation with a predefined number of clusters. For instance the local variable $x_1$, which corresponds to the height above the ground (in centimeters) is assigned to the nearest meters. We now define our latent CRF framework similar to [26, 25, 27, 28]. Consider a set of latent variable $\mathbf{h} = \{h_1, ..., h_M\}$, where each latent variable $h_i$ is obtained from observation $x_i$ given an arbitrary mapping function, We rewrite the conditional probability given in equation (2) as follows:

$$P(y, \mathbf{h}|\mathbf{x}; \Lambda) = \frac{1}{Z(\mathbf{x}; \Lambda)} \prod_{c \in C} \Psi_c(y, \mathbf{h}, \mathbf{x}; \Lambda) \tag{4}$$

Where, $\Lambda = \{\lambda^1, ..., \lambda^M\}$ and the partition function is now summed over all latent variables as well, i.e.

$$Z(\mathbf{x}; \Lambda) = \sum_{\mathbf{h}} \prod_{c \in C} \Psi_c(\mathbf{h}, \mathbf{x}; \Lambda)$$

The parameter set $\Lambda$ are estimated over the latent states, whereas $\theta$ in equation (2) corresponded to the parameters of feature functions over local observations $\mathbf{x}$. Summing the joint distribution given in equation (4) over all latent variables the conditional probability $P(y|\mathbf{x}; \Lambda)$ is obtained as:

$$P(y|\mathbf{x}; \Lambda) = \sum_h P(y, \mathbf{h}|\mathbf{x}; \Lambda) \tag{5}$$

The potentials $\Psi_c$ are computed differently depending on the feature functions of each clique. The feature functions $f_k^c$ are defined over $M$ different range of dependencies. The first order (unary potentials) are given by:

$$\Psi_{c_1}(y, \mathbf{h}, \mathbf{x}; \Lambda) = \exp\{\sum_{k \in k_{c_1}} \lambda_k^1 f_k^1(y, h_i, x_i)\} \tag{6}$$

where $k_{c_1}$ is a set of feature indexes for the unary parameters. The higher order feature functions are based on transition probabilities. There is one edge for every pairwise interactions and $\binom{M}{n}$ edges for interactions across the cliques

(a) (b)

**Figure 6:** Two examples of same body orientation (Frontal-View) with different visual appearances due to; a) different gimbal angles, and b) different headings.

of $n$ nodes, i.e.

$$\Psi_{c_n}(y, \mathbf{h}, \mathbf{x}; \Lambda) = \exp\{\sum_{k \in k_{c_n}} \lambda_k^n f_k^n(y, h_{1\ldots n}, x_{1\ldots n})\} \tag{7}$$

where, $k_{c_n}$ is a set of feature indexes of interaction parameters across $n$ nodes.

The main drawback of long range dependency analysis is that the CRF models are not easily scalable despite the fact that a particular set of states may highly be discriminative in deciding whether an object pattern exists or not. If the decision boundaries are too complex, the only way to discriminate the labels is to project the problem into a higher dimension where longer range of dependencies are considered. Let us consider an example to clarify the significance of some long range dependencies; Consider a UAV flight at 7 am, heading toward east at a GPS location close to earth's equator. In this scenario the sun is directly pointing at the camera and therefore the objects appears very dark (Figure 6). On the other hand, if the UAV is facing toward west in the same scenario the objects appear very bright and clear (Figure 6). Selecting the correct model (based on template) in this scenario without having prior knowledge of the scene is challenging. The time, heading, and location information alone are not the most discriminative information, however, when they are combined in one clique, the discriminative power of that clique increases significantly. In the next subsection we induce and evaluate all possible features during the training to search for such discriminative features.

### 3.4. Inducing Features based on Akaike Information Criterion

In section 3.3 we used latent conditional random fields (LCRF) to increase the sparsity of local observations and hence the expressive power of the features. However, the main question still remains, "Which features to pick?".

In order to answer this question we use the *corrected* Akaike Information Criterion (AICc) which promotes features that improve the conditional log-likelihood of the data and penalizes the features that are over fitting the model. The AICc definition is given as follows:

$$AICc = 2k - 2ln(L) + \frac{2k(k+1)}{n-k-1} \tag{8}$$

where $k$ is the number of parameters induced at each range of dependencies and $L$ is the maximum value of the likelihood given the corresponding feature functions. The underlying rational behind AICc is that if we knew the true distribution of the LCRF model, given in equation (4), we could calculate the exact information loss with measures such as Kullback-Leibler divergence. Nonetheless, AICc score measures the relative information gain before and after the feature is induced.

Similar to [24], each feature function $f_k^c$ is induced efficiently and the optimal parameters $\Lambda^*$ that maximizes the conditional log-likelihood of the data is estimated. The AICc is then evaluated with the induced feature and if the AICc score is reduced, the feature is selected, otherwise, it is discarded. The AICc score is asymptotically consistent, which means that the score is minimized for the graph structure $G$ as the number of samples grows to infinity [21]. The conditional likelihood $L$ is defined as $p(y|\mathbf{x}; \Lambda)$, which is obtained over the latent states as described in equation (4). In the next section, we describe a gradient ascent technique, where the likelihood $L$ is maximized with the optimal parameters $\Lambda^*$, such that $\Lambda^* = \underset{\Lambda}{\operatorname{argmax}} L(\Lambda)$.

### 3.5. Parameter Estimation and Inference

In order to estimate the parameters we follow the previous work on CRFs [22] and use the following objective function:

$$L(\Lambda) = \sum_{t=0}^{T} \log P(y^t|\mathbf{h}, \mathbf{x}^t; \Lambda) - \frac{1}{2\sigma^2}||\Lambda||^2$$

where the first term is the conditional log-likelihood of the data and the second term is the gaussian prior with variance $\sigma^2$. Hence, $P(\Lambda) = \frac{1}{2\sigma^2}||\Lambda||^2$. We estimate the parameters with the goal of maximizing the conditional likelihood term similar to [26, 25]. We use gradient ascent with BFGS optimization given that the likelihood term for the $t^{th}$ training sample is efficiently computed as follows:

$$L^t(\Lambda) = \log P(y^t|\mathbf{x}^t, \Lambda) = \log(\frac{\sum_{\mathbf{h}} \exp\{\Psi_c(y^t, \mathbf{h}, x_i, ; \Lambda)\}}{\sum_{y',h} \exp\{\Psi_c(y', \mathbf{h}, x^t; \Lambda)\}})$$

Taking the derivative of $L^t(\Lambda)$ with respect to the unary parameters we get:

$$\frac{\partial L^t(\Lambda))}{\partial \lambda_k^1} = \sum_{i,a} P(h_i = a|y^t, \mathbf{x}^t; \Lambda) f_k^1(i, y^t, a, \mathbf{x}^t) - \sum_{y',i,a} P(h_i = a, y'|\mathbf{x}^t; \Lambda) f_k^1(i, y', a, \mathbf{x}^t)\}$$

The marginal probability of edge features can also be efficiently computed with belief propagation, given that AICc criterion significantly reduces the number of long range features. Once the optimal parameters $\Lambda^*$ are known for each label (body orientation) the inference technique involves the following four steps: a) Computing the weighted sum of hidden state clique potentials for each training sample, which is proportional to $\sum_{\mathbf{h}:\forall h_i \in \mathcal{H}_i^y} P(\mathbf{h}|\mathbf{x};\Lambda)$; b) Computing the same weighted sum for the test instance (during real-time flight); c) Selecting the models $\mu^t$ that correspond to the K-nearest training samples; and d) Apply equation (1) to select the label $\hat{y}$ which has the highest visual similarity. In the experimental section we discuss the effect of choosing different $K$.

## 4. Experiments

In section 3 we proposed a new technique based on LDCRF to characterize the visual appearance of the detected object with the non-visual scene information obtained obtained from the UAV telemetry. Having learned a large number of visual appearance models, the objective is to leverage on scene characteristics to prune the search space of the models. In this section, we evaluate our method in the application of coronal plane estimation, where the task is to recognise the body orientation in a challenging outdoor environment. We first describe the UAV data acquisition procedure. We demonstrate how consistent patterns captured within the telemetry data can improve the visual model selection. The significance of each telemetry combination is evaluated in the LDCRF framework. Our experiments and results demonstrate thes advantages of the proposed approach compared to alternative model selection techniques such as CRF, LCRF, and Hamming Distance.

### 4.1. Dataset

Our dataset is collected over 124 flights. During each flight the UAV follows an octagonal trajectory around each person at different altitudes above the ground, gimbal angle, magnetic heading, location, and time of the Day. During data collection, the person is positioned in the center of a predefined octagon, and the UAV is programmed to follow a precise trajectory in order to obtain eight visual appearances of the person from each corner of the octagon at different telemetry settings. The viewpoint adjustments take place during the UAV transition from one point to another, and the frames collected during the transition are discarded. We use a GoPro camera and sample the images at 15 frame per second (fps) to match the frequency of telemetry readings. The telemetry measures are obtained with Kestrel III autopilot [1]. The full HD uncompressed videos (1920x1080 resolution) and five synchronized telemetry data are then transmitted to the ground with less than 2 ms latency using a low power (5V) PARALINX ARROW HD wireless transmitter which weighs less than 40 grams. The range of our HD wireless transmission is limited to 300ft line-of-sight which is well beyond the constraints posed in our problem statement.

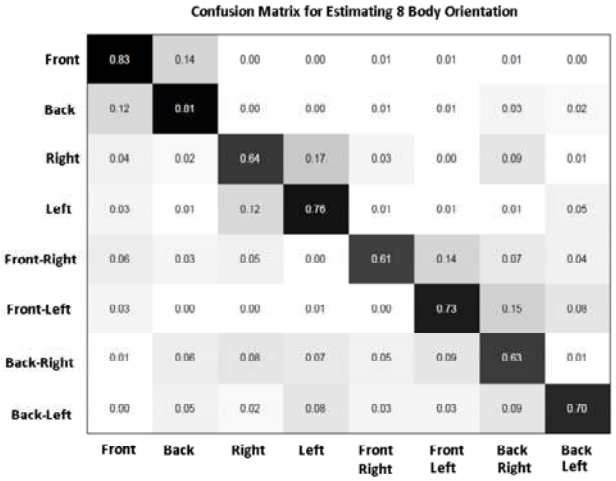


**Figure 7:** Confusion matrix for all body orientations using LDCRF technique. This result illustrate the consistent performance over all body orientations in variety of scene characteristics.

### 4.2. LDCRF over Telemetry Data

The mutual information between cliques of telemetry data is captured over five hidden states. The scope of each hidden state is as follow: **Altitude**$[2m - 9m]$ with increments of one meter; **Gimbal Angle** $[10° - 80°]$ with the increments of $10°$; **Heading** $[0° - 360°]$ with the increments of $45°$; **Location** $[1 - 5]$, where each location is separated by an average of two miles; and **Time** $[6am - 10am]$ with increments of 1 hour. Hence, for each body orientation we have 5 nodes and 26 edges $(\binom{5}{1} + \binom{5}{2} + \binom{5}{3} + \binom{5}{4} + \binom{5}{5})$. However, in order to fully exploit the mutual information, all possible node and edge feature functions needs to be evaluated, which could lead to a very large number (26,331 parameters which includes 34 possible cliques of one, 545 cliques of two, 3032 cliques of three, 9,920 cliques of four and 12,800 cliques of five). Therefore, during the training stage, we induce features one at a time and evaluate how well the model is fit, using AICc score, before including the feature in the model. This leads to more than $98\%$ rejected features. Averaged over eight labels, the mutual information in the telemetry data is modeled with 510 features. Figure 8 shows the minimization of negative log-likelihood with the selected features. The order in which the features of each clique are induced may slightly change this figure but the asymptotic decline stays more or less the same. The average performance shown in Figure 7.

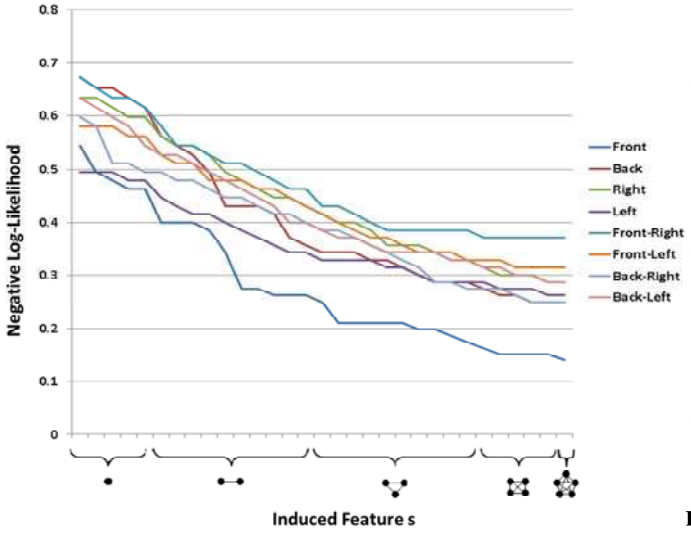


**Figure 8:** The negative log-likelihood (NLL) of the data is shown, where the features are incrementally added using AICc score. All possible features are induced and only the ones that minimizes the AICc score is added at each iteration. The order in wich the features are induced may result a slightly different graph but the asymptotic decline remains the same. This figure demonstrates the improvement made using higher order interactions as the NLL is minimized.

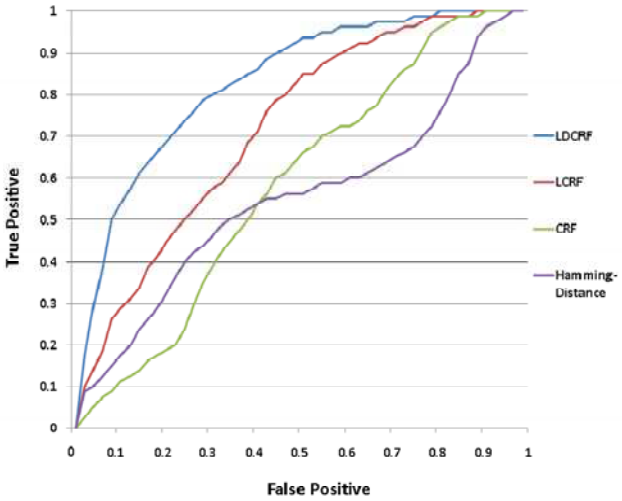


**Figure 9:** The ROC curve for detection of coronal plane is shown as a function of K-nearest models. LDCRF (depicted in blue) illustrates the best result due to consideration of highly discriminative features. LCRF (depicted in red) shows a worst performance due to a large number of redundant features. CRF performance (depicted in green) also performed poorly due to the densely distributed telemetry reading. Hamming distance performed worst than rest indicating the inadequacy of discriminative information in order to estimate the decision boundary.

### 4.3. Performance Comparison (CRF, LCRF, LDCRF, Hamming Distance)

Given that our main goal is to characterize the visual appearance with non-visual information obtained from the scene, any distance measure between the observed telemetry and the telemetry of available models can be used. We argue that such "distance metrics" require a consideration of higher order interactions. In order to illustrate the importance of higher order interactions we compare our results with the case where we used hamming distance between the telemetry of observed scene and the the associated telemetry of visual appearance models (the distance is computed over the hidden states). Figure 9 shows the ROC curve for the frontal face as a function of $K$, where $K$ is the number of selected nearest neighbors. Notice that the correct models are selected much faster (as a function of $K$) in the case of LDCRF. The performance drop in the case of LCRF is due to the noise induced by redundant features and the poor performance of CRF is due to the densely distributed features, which in turn makes the decision boundary more complex.

In the supplementary materials, we include both aerial and ground videos, where the orientation of the user is detected and displayed on every frame in real-time. The synchronized ground coverage illustrates the autonomous flight in which the UAV maneuvers to face the user once the coronal plane is estimated.

## 5. Conclusion

In this paper we introduced a new technique that links the non-visual scene characteristics to the visual appearance model of eight body orientations. We presented an autonomous UAV system capable of estimating the coronal plane of detected person followed by automatic maneuver to face the person. Proposed technique exploits the mutual information among telemetry, using latent-dynamic conditional random fields (LDCRF), to aid the classification problem with visual appearance models. Mutual information is evaluated over all ranges of dependency and only the most significant information is selected in the model using the Akaike Information Criterion (AIC). Experimental results illustrate the proposed method is fast (15 fps), the achieves an average accuracy of 72% over eight body orientations.


## Acknowledgements:

This research is sponsored by ONR Grant # N00014-12-1-0503 and by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation.

## References


[1] Kestrel Autopilot lockheedmartin. http://www.lockheedmartin.com/us/products/procerus/kestrel-autopilot.html. Accessed: 2010-09-30. 6

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998. 2, 3

[3] S. O. Ba and J.-M. Odobez. A rao-blackwellized mixed state particle filter for head pose tracking. In *ACM ICMI Workshop on Multimodal Multiparty Meeting Processing (MMMP)*, number EPFL-CONF-83238, 2005. 1

[4] B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *BMVC*, pages 1–11, 2009. 2

[5] B. Benfold and I. Reid. Unsupervised learning of a scene-specific coarse gaze estimator. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2344–2351. IEEE, 2011. 2

[6] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998. 2

[7] C. Chen, A. Heili, and J.-M. Odobez. A joint estimation of head and body orientation cues in surveillance video. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 860–867. IEEE, 2011. 2

[8] C. Chen and J.-M. Odobez. We are not contortionists: coupled adaptive learning for head and body orientation estimation in surveillance video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1544–1551. IEEE, 2012. 2

[9] R. T. Collins, Y. Liu, and M. Leordeanu. Online selection of discriminative tracking features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1631–1643, 2005. 2

[10] M. Enzweiler and D. M. Gavrila. Integrated pedestrian classification and orientation estimation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 982–989. IEEE, 2010. 2

[11] G. Fanelli, J. Gall, and L. Van Gool. Real time head pose estimation with random regression forests. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 617–624. IEEE, 2011. 1

[12] T. S. Farlow, M. T. Rosenstein, M. Halloran, C. Won, S. V. Shamlian, and M. Chiappetta. Mobile robot system, Feb. 19 2015. US Patent App. 14/625,646. 1

[13] F. Flohr, M. Dumitru-Guzu, J. F. Kooij, and D. M. Gavrila. Joint probabilistic pedestrian head and body orientation estimation. In *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, pages 617–622. IEEE, 2014. 2

[14] F. Flohr and D. M. Gavrila. Pedcut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues. In *Proc. BMVC*, pages 66–1, 2013. 2

[15] T. Gandhi et al. Image based estimation of pedestrian orientation for improving path prediction. In *Intelligent Vehicles Symposium, 2008 IEEE*, pages 506–511. IEEE, 2008. 2

[16] D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International journal of computer vision*, 73(1):41–59, 2007. 2

[17] H. Grabner and H. Bischof. On-line boosting and vision. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 260–267. IEEE, 2006. 2

[18] S. Hinterstoisser, O. Kutter, N. Navab, P. Fua, and V. Lepetit. Real-time learning of accurate patch rectification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2945–2952. IEEE, 2009. 3

[19] Z. Kalal, J. Matas, and K. Mikolajczyk. Pn learning: Bootstrapping binary classifiers by structural constraints. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 49–56. IEEE, 2010. 3

[20] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7):1409–1422, 2012. 2

[21] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 5

[22] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. 3, 5

[23] J. Lim, D. A. Ross, R.-S. Lin, and M.-H. Yang. Incremental learning for visual tracking. In *Advances in neural information processing systems*, pages 793–800, 2004. 2

[24] A. McCallum. Efficiently inducing features of conditional random fields. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 403–410. Morgan Kaufmann Publishers Inc., 2002. 5

[25] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 4, 5

[26] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):1848–1852, 2007. 4, 5

[27] A. Rahimi, R. J. Miller, D. Fedorov, S. Sunderrajan, B. Doheny, H. Page, and B. Manjunath. Marine biodiversity classification using dropout regularization. In *Computer Vision for Analysis of Underwater Imagery (CVAUI), 2014 ICPR Workshop on*, pages 80–87. IEEE, 2014. 4

[28] A. Rahimi, L. Nataraj, and B. Manjunath. Features we trust! In *IEEE International Conference on Image Processing (ICIP) 2015*. IEEE, 2015. 4

[29] A. Schulz, N. Damer, M. Fischer, and R. Stiefelhagen. Combined head localization and head pose estimation for video–based advanced driver assistance systems. In *Pattern Recognition*, pages 51–60. Springer, 2011. 2

[30] H. Shimizu and T. Poggio. Direction estimation of pedestrian from multiple still images. In *Intelligent Vehicles Symposium, 2004 IEEE*, pages 596–600. IEEE, 2004. 2