2009 IEEE International Conference on Robotics and Automation
Kobe International Conference Center
Kobe, Japan, May 12-17, 2009

Expansion Segmentation for Visual Collision Detection and Estimation

Jeffrey Byrne and Camillo J. Taylor

Abstract— **Collision detection and estimation from a monocular visual sensor is an important enabling technology for safe navigation of small or micro air vehicles in near earth flight. In this paper, we introduce a new approach called *expansion segmentation*, which simultaneously detects "collision danger regions" of significant positive divergence in inertial aided video, and estimates maximum likelihood time to collision (TTC) in a correspondenceless framework within the danger regions. This approach was motivated from a literature review which showed that existing approaches make strong assumptions about scene structure or camera motion, or pose collision detection without determining obstacle boundaries, both of which limit the operational envelope of a deployable system. Expansion segmentation is based on a new formulation of 6-DOF inertial aided TTC estimation, and a new derivation of a first order TTC uncertainty model due to subpixel quantization error and epipolar geometry uncertainty. Proof of concept results are shown in a custom designed urban flight simulator and on operational flight data from a small air vehicle.**

I. INTRODUCTION

Safe and routine operation of autonomous vehicles requires the robust detection of hazards in the path of the vehicle, such that these hazards can be safely avoided without causing harm to the vehicle, other objects or bystanders. Obstacle detection approaches have been successfully demonstrated on autonomous ground vehicles, notably in the DARPA grand challenge events, including extended collision free operation in both off-road and controlled urban terrain. These vehicles have sufficient size, weight and power (SWAP) capabilities to support active sensors such as LIDAR or millimeter wave RADAR, or use of a dominant ground plane to aid in visual obstacle detection.

In contrast, small or micro air vehicles (MAVs) are small, lightweight, and autonomous aerial systems that can fit in a backpack, and promise to enable on-demand intelligence, surveillance and reconnaissance tasks in a near-earth environment. To move towards routine MAV flight in a near earth environment, we first must demonstrate an "equivalent level of safety" [1] to a human pilot using appropriate sensors for the platform. Unlike ground vehicles, MAVs introduce aggressive maneuvers which couples full 6-DOF platform motion with sensor measurements, and feature significant SWAP constraints that limit the use of active sensors. Even those active sensors that have potential for deployment on small UAVs [2] take away SWAP required for the payload to achieve the primary mission, and such approaches will not scale to the smallest MAVs. Furthermore, the wingspan limitations of MAVs limit the range resolution of stereo configurations [3], therefore an appropriate sensor for collision detection on a MAV is monocular vision. While monocular collision detection has been demonstrated in controlled flight environments [4][5][6][7][8][9][10][11][12], it remains a challenging problem due to the low false alarm rate needed for practical deployment and the high detection rate requirements for safety. In this paper, we review the literature on current approaches for visual collision detection and estimation, and propose a new method to reflect this analysis called *expansion segmentation*. This method combines visual collision detection to localize significant collision danger regions in forward looking aerial video, with optimized time to collision estimation within the collision danger region. Formally, expansion segmentation is the labeling of "collision" and "non-collision" nodes in a conditional Markov random field. The minimum energy binary labeling is determined in an expectation-maximization framework to iteratively estimate labeling using the min-cut of an appropriately constructed affinity graph, and the parameterization of the joint probability distribution for time to collision and appearance. This joint probability provides a global model of the collision region, which can be used to estimate maximum likelihood time to collision over optical flow likelihoods, which is used to aid with local motion correspondence ambiguity.

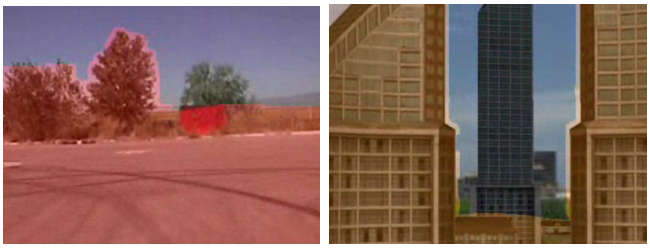

Fig. 1. Expansion Segmentation provides (i) detection of significant "collision dangers" in inertial aided video shown as a semitransparent overlay and (ii) mean maximum likelihood time to collision estimation in seconds within the danger region shown by color (yellow=far, red=close)

New contributions of this work are as follows:

- Expansion segmentation theory and experimental results as a new approach simultaneous collision detection and estimation in a correspondenceless framework.
- Derivation of visual time to collision estimation using inertial aiding

This work was supported by an NDSEG graduate research fellowship and AFRL/MNGI contract FA8651-07-C-0094

J. Byrne has a joint appointment with GRASP Lab, Department of Computer and Information Science, University of Pennsylvania and Scientific Systems Company (SSCI) jebyrne@cis.upenn.edu

C.J. Taylor is an associate professor with the GRASP Lab, Department of Computer and Information Science, University of Pennsylvania cjtaylor@cis.upenn.edu

- Derivation of a time to collision uncertainty model showing inertial aiding is crucial to detect small obstacles in urban flight
- Explicit use of derived time to collision uncertainty model within the expansion segmentation framework.
- Custom designed urban flight simulator with ground truth for closed loop performance evaluation

II. RELATED WORK

The dominant approaches in the literature for monocular visual collision detection and estimation can be summarized in four categories: structure from motion, ground plane methods, flow divergence and insect inspired methods.

Structure from motion (SFM) is the problem of recovering the motion of the camera and the structure of the scene from images generated by a moving camera. SFM techniques [13] provide a sparse or dense 3D reconstruction of the scene up to an unknown scale and rigid transformation, which can be used for obstacle detection when combined with an independent scale estimate for metric reconstruction, such as from inertial navigation to provide camera motion or from a known scene scale. Modern structure from motion techniques generate impressive results for both online sequential and offline batch large scale outdoor reconstruction. Recent applications relevant to this investigation include online sparse reconstruction during MAV flight for downward looking cameras [14], and visual landing of helicopters [15][16]. However, SFM techniques consider motion along the camera's optical axis as found in a collision scenario to be degenerate due to the small baseline, which results in significant triangulation uncertainty near the focus of expansion (see section III) which must be modelled appropriately for usable measurements.

Ground plane methods [17][18], also known as horopter stereo, stereo homography, ground plane stereo or inverse perspective mapping use an homography induced by a known ground plane, such that any deviation from the ground plane assumption in an image sequence is detected as an obstacle. This approach has been widely used in environments that exhibit a dominant ground plane, such as in the highway or indoor ground vehicle community, however the ground plane assumption is not relevant for aerial vehicles.

Flow divergence methods rely on the observation that objects on a collision course with a monocular image sensor exhibit expansion or *looming*, such that an obstacle projection grows larger on the sensor as the collision distance closes [19][20]. This expansion is reflected in differential properties of the optical flow field, and is centered at the focus of expansion (FOE). The FOE is a stationary point in the image such that expansion rate from the FOE or *positive divergence* is proportional to the time to collision. Flow divergence estimation can noisy due to local flow correspondence errors and the amplifying effect of differentiation, so techniques rely on various assumptions to improve estimation accuracy. These include assuming a linear flow field due to narrow field of view during terminal approach [19][21][20][22], assuming known camera motion and positioning the FOE at image center [23][24][25][26],

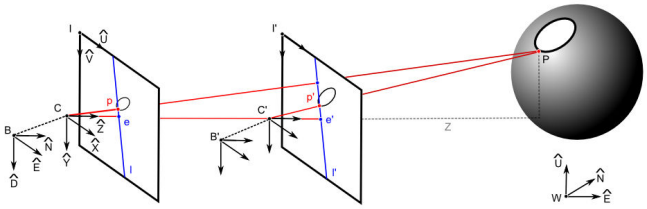

Fig. 2. Epipolar geometry for time to collision estimation

or known obstacle boundaries for measurement integration [22][23][27]. These strong assumptions limit the operational envelope, which have lead some researchers to consider the qualitative properties of the motion field rather than metric properties from full 3D reconstruction as sufficient for collision detection [28][20]. However, this does not provide a measurement of time to collision and does not localize collision obstacles in the field of view.

Insect vision research on the fly, locust and honeybee show that these insects use differential patterns in the optical flow field to navigate in the world. Specifically, research has shown that locusts use expansion of the flow field or "looming cue" to detect collisions and trigger a jumping response [29]. This research has focused on biophysical models of the Lobula Giant Movement Detector (LGMD), a wide-field visual neuron that responds preferentially to the looming visual stimuli that is present in impending collisions. Models of the LGMD neuron have been proposed [30] which rely on a "critical race" in an array of photoreceptors between excitation due to changing illumination on photoreceptors, lateral inhibition and feedforward inhibition, to generate a response increasing with photoreceptor edge velocity. Analysis of the mathematical model underlying this neural network shows that the computation being performed is visual field integration of divergence for collision detection, which is tightly coupled with motor neurons to trigger a flight response. This shows that insects perform collision detection, not reconstruction. This model has been implemented on ground robots for experimental validation [12][31][32], however the biophysical LGMD neural network model has been criticized for lack of experimental validation [33], and robotic experiments have shown results that do not currently live up to the robustness of insect vision, requiring significant parameter optimization and additional flow aggregation schemes for false alarm reduction [34][35].

III. INERTIAL AIDED TIME TO COLLISION

In this section, we formulate the problem of estimating time to collision using inertial aiding, and provide an uncertainty analysis for this estimate.

A. Inertial Aided Epipolar Geometry

Figure 2 shows a calibrated camera C rigidly mounted to a body frame B moving with a translational velocity V and rotational velocity Ω. The body frame moves from B to B', and the camera captures perspective projections

I and I' at a sampling rate t_s of 3D point P in camera frames C and C' respectively. The camera C is intrinsically calibrated (K), the images (I) are lens distortion corrected, and the rotational alignment from to body to the camera ${}^{C}_{B}R$ is known from extrinsic calibration. The body orientation ${}^{B}_{W}R$ and position ${}^{B}_{W}t$ is estimated at B and B' relative to an inertial frame W from an inertial navigation system. Using Craig notation [36], the relative transform between camera frames from C to C' is ${}^{C'}_{C}T = ({}^{C'}_{B'}T\,{}^{B'}_{W}T)({}^{C}_{B}T\,{}^{B}_{W}T)^{-1}$, where ${}^{C'}_{C}R$ is the upper 3x3 submatrix of ${}^{C'}_{C}T$. Define a rotational homography $H = K({}^{C'}_{C}R)K^{-1}$, and the projection matrix $({}^{C}_{W}P)$ which is the upper 3x4 submatrix of $({}^{C}_{W}T) = ({}^{C'}_{B'}T\,{}^{B'}_{W}T)$, then the focus of expansion or epipole $e = K({}^{C}_{W}P)({}^{B'}_{W}t)$ which is the projection of the origin of C' in C. Given an estimate of the essential matrix $E = {}^{C'}_{C}\hat{T}\,{}^{C'}_{C}R$ from inertial aided epipolar geometry, compute the epipolar line $l' = K^{-T}EK^{-1}p$, such that corresponding points p and p' which are constrained to fall on epipolar lines l and l'. Finally, the time to collision (τ') relative to C' to P is:

$$\tau' = \frac{Z}{V} = \frac{(p-e)^T(p-e)}{(p-e)^T(p'-Hp)}t_s \tag{1}$$

where the rotation compensating homography H and epipole e are determined from inertial aiding.

Intuitively, the time to collision τ' is determined by the distance of a point p from the epipole divided by the rate of expansion from the epipole due to translation only, with rotational effects removed. τ' is completely determined from image correspondences p and p' as well as inertial aided measurements H, e and sampling rate t_s. Note that in this formulation, "collision" is defined as the time required for point P to intersect with an *infinite image plane* at instantaneous velocity V, which depending on the extent of the vehicle body may or may not pose an immediate collision danger on the current trajectory. The full derivation of equation (1) follows directly from the motion field, with rotational homography and epipole assumed known from inertial aiding.

B. Time to Collision Uncertainty Analysis

Without loss of generality, define the epipole e to be at the image origin, such that equation (1) simplifies to $\tau = p/\dot{p}$, where p is the Euclidean distance from the origin, and $\dot{p} = v$ is the radial rate of expansion along epipolar lines due to translation only. Model p as a Gaussian random variable with parameterization $N(\mu_p, \sigma_p^2)$, such that the variance σ_p^2 is determined from the expected subpixel accuracy of p. Model v as a difference two Gaussian random variables p' and p, forming a discrete approximation to the temporal derivative. Assuming independent measurements, a difference of Gaussians can be modeled with parameterization $N(\mu_v, \sigma_v^2) = N(\mu_{p'} - \mu_p, 2\sigma_p^2)$.

Consider a first order Taylor series expansion of τ which is a function $\tau(p,v)$ about the point (μ_p, μ_v).

$$\tau \approx \tau(\mu_p,\mu_v) + (p-\mu_p)\frac{\partial\tau(\mu_p,\mu_v)}{\partial p} + (v-\mu_v)\frac{\partial\tau(\mu_p,\mu_v)}{\partial v} \tag{2}$$

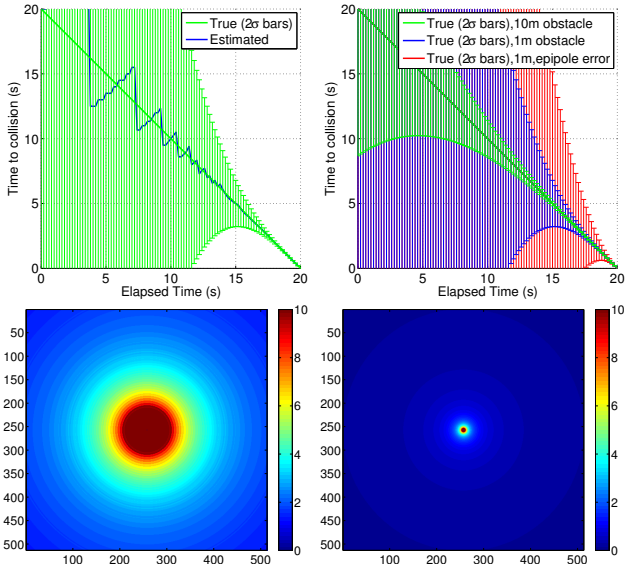

Fig. 3. (top) Time to collision theoretical uncertainty. (bottom) Standard deviation of time to collision measurements of an obstacle at 200m and 20m as a function of image position.

The variance σ_τ^2 of the time to collision about the point (μ_p, μ_v) is given by the expectation

$$\sigma_\tau^2 = E\left[(\tau - \tau(\mu_p,\mu_v))^2\right] \tag{3}$$

Simplifying (3) using the Taylor series approximation in (2) results in

$$\sigma_\tau^2 = \frac{\mu_v^2\sigma_p^2 + \mu_p^2\sigma_v^2}{\mu_v^4} \tag{4}$$

Equation 4 is the uncertainty for a single point projection p, due to subpixel pixel quantization error. (4) is a first order approximation for the time to collision variance in terms of the Gaussian parameterization of position and expansion measurements. This variance estimate does not imply that τ is Gaussian. In fact, τ follows a *ratio distribution*, for which the variance approximation should be interpreted as a guide for the relative accuracy of time to collision measurements as determined from the second moment of a ratio distribution, rather than providing any probabilistic guarantees.

The time to collision uncertainty in (4) can also be due to epipolar geometry errors in addition to pixel quantization errors. This error is dominated by errors in the epipole location, however since the derivation assumes without loss of generality that the epipole is at the origin, epipole errors are modelled as appropriate increases of σ_p and σ_v.

Figure 3 (top left) shows an example of the time to collision uncertainty model in (4). In this example, a camera is moving at constant velocity along the optical axis such that it will collide with an obstacle in 20 seconds. The green plot shows the true (linear) time to collision along with 2σ uncertainty as determined from (4) for a fixed point on the obstacle 1m orthogonal to the optical axis. The blue curve shows the estimated time to collision assuming 0.25 subpixel interpolation accuracy and focal length f=1000 pixels. Notice

that the estimate exhibits a characteristic "staircase" pattern, which is due to the pixel quantization for p changing faster than $\dot{p}$ at large TTC, however the effects of quantization are reduced as the collision distance closes. Figure 3 (bottom) shows the standard deviation from (4) as a function of image position, which shows that for an obstacle at constant distance, the uncertainty significantly increases nearer to the focus of expansion and for closer obstacles. Finally, figure 3 (top right) shows three time to collision uncertainty plots for a 10m obstacle, 1m obstacle and 1m obstacle with uncertainty in epipolar geometry. Urban obstacles such as traffic lights, poles, and signs (not including wires) are commonly of the order of 1m the largest dimension. This plot shows that the uncertainty model down to 1m obstacles are reasonably accurate at approximately 7s to collision. However, if the epipolar geometry is determined from online egomotion estimates rather than inertial aiding, then the location of the epipole may deviate (in our experience) by approximately 0.5^o CEP.

From this analysis, we draw two conclusions. First, inertial aiding is crucial for practical urban flight which may contain objects smaller than 1m. Second, TTC exhibits an anisotropic uncertainty based on image position as shown in figure 3 (bottom), and the TTC estimates are sensitive to subpixel correspondence errors at larger standoff distances. Therefore, due to the magnitude of these errors, they must be appropriately modelled during time to collision estimation to achieve accuracy necessary for safe flight.

IV. EXPANSION SEGMENTATION

Expansion segmentation is a new approach to visual collision detection to find dangerous collision regions in inertial aided video while optimizing time to collision estimation within these regions. More formally, expansion segmentation is a grouping of pixels into collision and non-collision regions using joint probabilities of expanding motion and color, determined from a minimum energy binary labeling of collision and non-collision of a conditional Markov random field in an expectation-maximization framework.

Expansion segmentation (ES) addresses the key observations determined from the literature review of section II and the uncertainty analysis in section III-B. First, this method provides both collision detection and estimation, where the detection provides an aggregation or grouping of all significant expansion in an image. This approach does not assume known structure or known obstacle boundaries. Second, this method handles the geometric time to collision uncertainty discussed in section III-B by incorporating the uncertainty model into the detection and estimation framework. Third, this method handles sensitivity to local correspondence errors by using motion correspondence *likelihoods* rather than discrete correspondences. The global joint probability of time to collision and color for the detected danger region is used to aid in local correspondence. This approach is a *correspondenceless* method, as it does not rely on a priori correspondences as input. Our approach is inspired by [3][37][38][39][40][41][42][43], but it deviates from the literature in the explicit use of time to collision uncertainty model during labeling and region parameterization, and in the use of correspondenceless motion likelihoods.

Given two images I and I' with epipolar geometry H and e as determined from inertial aiding, expansion segmentation is a minimum energy solution to

$$E(f,\theta) = \sum_{i \in I} D(f_i, \theta; H, e, \delta_i, \tau_c, t_s) + \sum_{(i,j) \in \mathcal{N}} V(f_i, f_j; \gamma) \quad (5)$$

over both binary labels $f_i \in \{0,1\}$ for each of N pixels resulting in an image labeling $f = \{f_0, f_1, ...f_N\}$ in I. The labeling $f_i = 0$ corresponds to "collision", and $f_i = 1$ to "non-collision". $\theta = \{\theta_c, \theta_s\}$ is a global parameterization for joint probability of collision labeled features (θ_c) and non-collision labeled or "safe" features (θ_s). These joint probability distributions are defined over image feature measurements z modelled as a mixture of Gaussians, such that for all measurements z_i with label $f_i = 0$:

$$p(z|\theta_c) = \sum_i \alpha_i exp(-(z_i - \mu_i)^T \Sigma_i^{-1} (z_i - \mu_i)) \quad (6)$$

where α_i are normalized mixture coefficients and $\theta_c = \{\mu_1, \Sigma_1, ..., \mu_k, \Sigma_k\}$ is a parameterization for a mixture of k Gaussians of the joint distribution of image measurements z which have label 0 ("collision"). $p(z|\theta_s)$ is defined similarly for measurements with label 1 ("safe"). The number k is determined by the total number of measurements in an overcomplete manner. This global model makes the strong assumption that given the current image, measurements (e.g. TTC and color) are *correlated*, and this correlation is reflected in the joint and can be used to resolve local correspondence ambiguities. This assumption does not hold in general, and can result in errors, however there is a fundamental tradeoff between the complexity of the global model and the promise of real time performance.

D in (5) is the data term which encodes the cost of assigning label "collision" or "non-collision" f_i to $i \in I$, given global parameterization of the joint distribution of collision feature measurements θ_c and non-collision θ_s. This data term requires the following additional fixed inputs: (i) H and e which are the rotational homography and epipole from inertial aided epipolar geometry as discussed in section III-A, (ii) τ_c which is a threshold set by the operator which characterizes the time to collision at which an obstacle exhibits an *operationally relevant* risk, such that $\tau \leq \tau_c$ exhibits "significant" collision danger given the constraints of the vehicle and mission, (iii) t_s is the sampling rate of images I and I' for unit conversion of frames to collision to seconds to collision and (iv) $\delta_i(i')$ is a correspondence likelihood function between pixels $i \in I$ and $i' \in I'$, such that the maximum likelihood correspondence for i is $j^* = argmax_j\ \delta_i(j)$, with correspondence likelihood δ_i^*. This function provides a motion likelihood for each pixel i, and may use inertial aided epipolar geometry to limit the domain of δ_i. Experimental details of this function are provided in section V-A.

D in (5) captures the cost of assigning collision labels to a pixel i given image feature measurements. These measurements include a scalar estimate of time to collision

given $\delta_i(i')$ with $\tau_i(i')$ from equation (1), and 3 luminance and chrominance components of color c. The result is a measurement vector $z_i = [\tau\ c]$, for which we define two probability distributions as weighted integrals for each i:

$$P(\tau_i \leq \tau_c|\theta_c) = \max_j \delta_i(j) \int_{-\infty}^{\infty} p(z|\theta_c)N(\mu_i,\Sigma_i)dz \quad (7)$$

and $P(\tau_i > \tau_c|\theta_s)$ respectively. This models the probability that $\tau_i \leq \tau_c$ by integrating the joint PDF $p(z|\theta_c)$ from (6) over a Gaussian model of uncertainty of z_i, where $\mu_i = [\tau_i\ c_i]$ and $\Sigma_i = diag(\sigma_\tau, \sigma_c)$. τ_i is determined from eq (1) and σ_τ from eq (4). The result is a likelihood that the time to collision τ_i for the i^{th} pixel is "significant" (e.g. $< \tau_c$) using the derived uncertainty model for time to collision from section III-B. Finally, the data term D in (5) takes the form for binary labels f:

$$D = (1-f_i)P(\tau_i \leq \tau_c|\theta_c) + (f_i)P(\tau_i > \tau_c|\theta_s) \quad (8)$$

Equation (7) which models TTC uncertainty for the data likelihood in (8) using motion likelihoods δ_i in a correspondenceless framework is a central contribution of this work.

V in (5) is a function which encodes the cost of assigning labels f_i to i and f_j to j when (i,j) are neighbors in a given neighborhood set $\mathcal{N} \subset I \times I'$. This function represents a penalty for violating label smoothness for neighboring (i,j). In this formulation, the interaction term V takes the form of a Potts energy model with static cues based on the appearance measurement in the current image [44], forming a conditional random field:

$$V(f_i,f_j) = \gamma T(f_i \neq f_j) exp(-\beta|I(i)-I(j)|^2) \quad (9)$$

where T is 1 if the argument is true, and zero otherwise. This term will bias the labeling towards smooth labeling, with label discontinuities at edges with color differences. γ is a *smoothness parameter* which will encode the strength of the smoothness prior, and β is a measurement variance for color differences. Experiments in [42] show that the segmentation is insensitive to the choice of γ and for 4-neighbor connectivity, and a choice of $\gamma = 25$ provides stable segmentations across a range of scenes.

The minimization of (5) can be performed in an expectation-maximization (EM) framework to iteratively estimate the optimal labeling f given region parameterization θ (maximization), followed by an estimate of the maximum likelihood region parameterization given the labeling (expectation). The region parameterization θ is initialized to either a uniform distribution or set to the parameterization determined from the prior segmentation result. Given θ, the labeling in equation (5) can be solved exactly for a *binary labeling* by posing a maximum network flow problem on a specially constructed network flow graph which encodes (5) [45][46], for which efficient maxflow solutions are available [40]. Then, given this labeling, the region parameterizations θ_c and θ_s can be updated using only measurements z_i with labels $f = 0$ and $f = 1$ respectively. The Gaussian mixture parameters in (6) are exactly $\mu_i = [\tau_i\ c_i]$ and $\Sigma_i = diag(\sigma_\tau, \sigma_c)$ from equation (7), with mixture coefficients $\alpha_i = \delta_i^*$. This mixture takes into account the correspondence likelihood and uncertainty of τ_i based on the image position i.

Following convergence of the EM iteration, such that the labeling does not change significantly or a maximum number of iterations is reached, the output of expansion segmentation is the final labeling f^* such that labels $f_i = 0$ are "significant collision dangers" and the final collision region parameterization θ_c^*. The maximum likelihood time to collision for measurements within the collision danger region (all i labeled $f_i = 0$) can be estimated using (θ_c^*) as follows:

$$\tau_i^* = \underset{j}{\text{argmax}} P(\tau_i(j) \leq \tau_c|\theta_c^*) \quad (10)$$

for which $\tau_i(j)$ is determined from equation (1) such that correspondence (i,j) determines $\dot{p}$. This estimate uses the joint θ_c^* to estimate the maximum likelihood τ_i given the uncertainty model of time to collision, which provides global region information to optimize over the local correspondence likelihood function δ_i.

V. RESULTS

A. Experimental Setup

Video and inertial flight data were collected by flying a Kevlar reinforced Zagi fixed wing air vehicle in near earth collision scenarios with an analog NTSC video transmitter and a Kestrel autopilot with MEMS grade IMU wirelessly downlinked to a ground control station for video and telemetry data collection. Example imagery collected is shown in figure 4 (bottom row).

Urban flight data collection is infeasible due to regulatory constraints of urban flight and the challenge of collecting dense ground truth. Instead, we created a custom flight simulation environment based on Matlab/Simulink and OpenSceneGraph in which to test algorithms for closed loop visual collision detection, mapping and avoidance. This provides medium fidelity rendered video of 3D models and terrain in "Megacity", ground truth range for performance evaluation, and a validated model of inertial navigation system measurements for inertial aiding. Example imagery from Megacity are shown in figure 4. The ground truth range to obstacles is not shown, but is used for quantitative performance evaluation.

The experimental system to test expansion segmentation implemented the following processing chain:

1) Bouguet intrinsic camera calibration and Lobo inertial-camera extrinsic calibration [47]
2) Preprocessing for video deinterlacing and RGB to YUV color space conversion
3) Analog video noise classification to classify noisy frames during downlink from the air vehicle [48]
4) Scaled and oriented feature extraction using steerable filters [49][50]
5) Motion likelihood from steerable filter phase correlation [51] with inertial aiding in a correspondenceless framework
6) Expansion segmentation with maximum likelihood time to collision estimation (section IV)

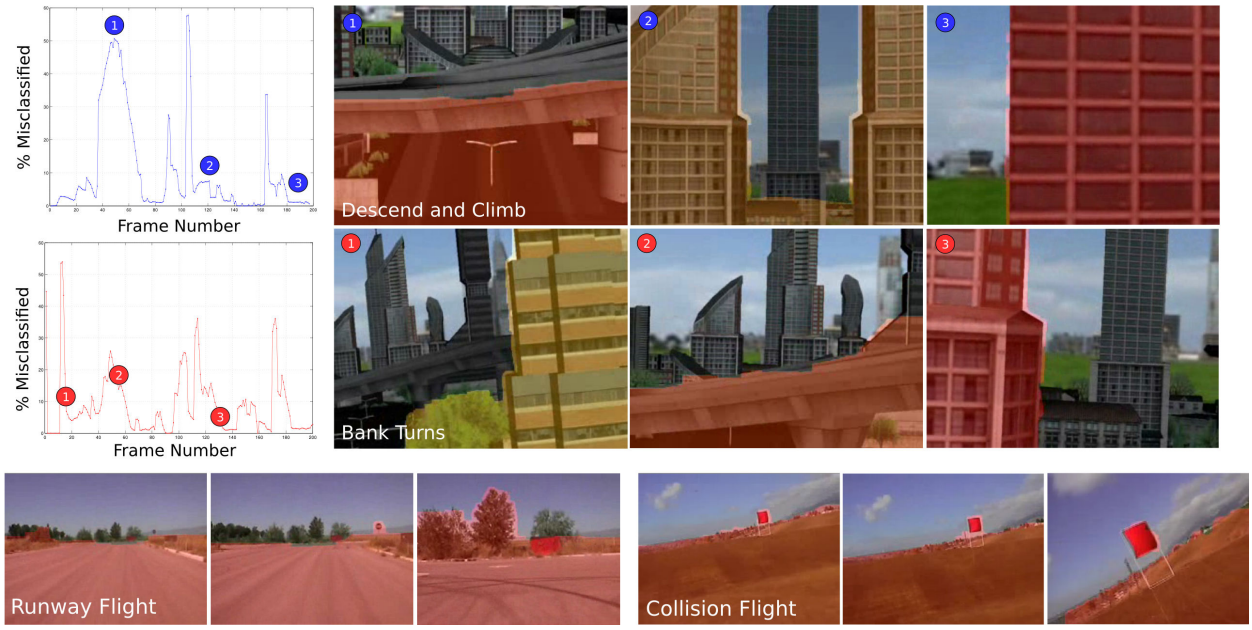

Fig. 4. Proof of concept expansion segmentation results. Collision detection shown as semi-transparent overlays with yellow, orange to red color encoding the time to collision estimate. (top) Descend and climb performance in Megacity (middle) Bank turn performance in Megacity (bottom row) Qualitative expansion segmentation results on operational video and telemetry. See the associated video for additional results.

The motion likelihood in step 5 is the implementation of δ_i in eq 5. This approach uses phase correlation of quadrature steerable filter responses of two images I and I' [51], using inertial aiding to provide epipolar lines as constraints for correspondence. Phase correlation is implemented as a disparity likelihood within a fixed disparity range (d_{max}) and orthogonal distance threshold (ρ_{max}) from epipolar lines. The orthogonal epipolar projection length ρ of p' onto the epipolar line l' is

$$\rho^2 = \frac{\left(p'^T F p\right)^2}{\|\hat{e}_3 F p\|^2} \tag{11}$$

ρ_{max} is chosen experimentally to reflect the uncertainty in the inertial aided epipolar geometry, and d_{max} is chosen relative to τ_c. Phase correlation is computed for all epipolar inliers p' using bilinear interpolation of features at integer disparity along epipolar lines. In eq 11, $\hat{e}_3$ is the cross product matrix for $e_3 = [0\ 0\ 1]^T$ and F is the fundamental matrix where $F = K^{-T}EK^{-1}$. The result is a motion likelihood function $\delta_p(p')$ as determined from phase correlation *over all inliers* (p').

Experiments with the Kestrel autopilot and MEMS grade IMU showed that the rotational homography H can be directly computed from inertial measurements, however position errors due to accelerometer biases and GPS uncertainties contribute significant error to the epipolar geometry. In this experimental system, we use a random sample of SIFT feature correspondences and sparse bundle adjustment [52] initialized with the inertial measurement to improve the essential matrix estimate [53].

All results in this section were generated using the following parameters: 320x240 imagery, 9x9 steerable filter kernels, $\mathcal{N}$ is 4-neighbor connectivity, $\rho_{max} = 0.5$, $d_{max} = 24$, 0.5 subpixel disparity, γ=25, θ is initialized to uniform distribution, and θ in (7) is implemented as a joint histogram with fixed bin width rather than mixture of Gaussians. In our experience, this is an suitable approximation which does not significantly impact performance. The experimental system is implemented in C++ with Matlab MEX wrappers for data visualization, and converges in 5-12 EM iterations in approximately 5 seconds per image on a 2.2GHz Intel Core 2 Duo. In our benchmarks, δ_p computation of motion likelihood dominates runtime performance and can be further optimized.

B. Simulation Results

Figure 4 shows expansion segmentation results on simulated and operational flight data. Figure 4 (top) shows quantitative performance evaluation of a descend and climb scenario in the Megacity simulation environment. The percent misclassification is the percentage of pixels incorrectly classified as either dangerous (false positive) or safe (missed detection) for a $\tau_c = 10s$ relative to the ground truth. This performance metric is widely used in the evaluation of stereo algorithms [54] and is adapted here for evaluation of time to collision. Expansion segmentation results are shown at three points in the scenario, where the color of the semi-transparent overlay encodes the mean time to collision for the danger region (yellow=far, red=close). The large percentage misclassification at (1) is due to the classification of the road underneath the overpass as dangerous, as it has few strong features for feature correspondence. The

misclassification at (2) is due pixels at the border having no motion measurement resulting in a smoothing of the image border into the foreground. Figure 4 (middle) shows a bank turn scenario in Megacity with misclassifications due to smoothing at the image border. In both scenarios, large narrow spikes in misclassification are due to the expansion segmentation not yet detecting that a large foreground region is dangerous due to time to collision uncertainty. Smaller misclassifications are due to motion ambiguity from periodic features, oversmoothing at the image edges where there are no motion measurements and time to collision uncertainty near the epipole.

All results are best viewed in the associated video, which also includes additional results not shown in figure 4. The video shows an expanding central square in a uniform random and sparse binary image, which demonstrates that motion only without any color information can be used for collision detection, and smoothness in the conditional random field can be used to "fill in the gaps" in the sparse binary image rather than detecting the white dots only. Next, the video shows a gray square offset from the checkerboard background by a fixed distance such that the gray square is expanding at a faster rate than the background. In this example, the joint region parameterization θ for color and motion is used to successfully segment the interior of the gray square which has no contrast for feature correspondence. The remainder of the video shows qualitative expansion segmentation videos of the scenarios described above.

C. Flight results

Figure 4 (bottom) shows qualitative results for operational flight data. First, data was collected on a runway during takeoff, and results show that the road, trees, fence and red tarp all exhibit a significant collision danger while the central tree and right mountains are set back in the scene and therefore do not exhibit immediate collision danger and are correctly detected as "safe". Note that collision dangers are defined as the time to intersect an infinite image plane, so peripheral trees and stop sign are correctly detected as potential collisions. Also, note that at no time is a ground plane assumption used to generate these results, and for an aerial vehicle the ground is a legitimate collision danger. The time to collision for these regions is dominated by the ground plane which has a small time to collision to intersect the infinite image plane, so therefore the color of the semi-transparent overlay is consistently red. The video shows time to collision with $\tau_c = 5s$ and $\tau_c = 8s$ showing that the trees are detected earlier for $\tau_c = 8s$. Quantitative evaluation was not performed due to a lack of ground truth for the flight sequences.

Finally, data was collected during a true collision event of a single high contrast obstacle with a human pilot in the loop for safety. The expansion segmentation results are best viewed in color and magnified in the PDF or in the associated video. This result shows that the collision danger regions are successfully segmented in full 6-DOF motion from a small UAV, and thus demonstrating proof of concept.

VI. CONCLUSIONS AND FUTURE WORK

The results demonstrated in simulation and on flight data are preliminary and show results only on a limited dataset, however they successfully prove the concept that expansion segmentation is feasible approach for visual collision detection and estimation on operational data from a small air vehicle. Additional test and evaluation is needed including urban driving as a surrogate for flight data and extended urban simulations, along with quantitative performance evaluation on long runs to determine false alarm and detection rates. Furthermore, comparative results are needed to show performance relative to other methods such as the approach described [53], however the lack of a standard inertial aided dataset makes direct comparisons difficult.

Finally, the expansion segmentation approach is promising for other applications which may be explored, including target pursuit which requires nulling the effects of expansion, and expansion segmentation due to zoom for foreground/background segmentation.

VII. ACKNOWLEDGMENTS

The authors gratefully acknowledge the help of Jeff Saunders and members of the MAGICC lab at Brigham Young University, who provided the flight data collection, and Benjamin Cohen at University of Pennsylvania who provided support for camera calibration and data analysis.

REFERENCES

[1] D. M. et. al, "Regulatory and technology survey of sense-and-avoid for uas," in *Proceedings AIAA Infotech@Aerospace*, May 2007.

[2] L. C. S. Scherer, S. Singh and S. Saripalli, "Flying fast and low among obstacles," in *Proceedings International Conference on Robotics and Automation*, April 2007.

[3] J. Byrne, M. Cosgrove, and R. Mehra, "Stereo based obstacle detection for an unmanned air vehicle," in *IEEE International Conference on Robotics and Automation (ICRA'06)*, May 2006.

[4] G. Barrows, "Mixed mode vlsi optic flow sensors for mavs," Ph.D. dissertation, University of Maryland at College Park, 1999.

[5] T.Netter and N.Franceschini, "A robotic aircraft that follows terrain using a neuromorphic eye," in *Proc. 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2002)*, 2002.

[6] C. M. Higgins, "Airborne visual navigation using biomimetic vlsi vision chips," Higgins Laboratory, University of Arizona Department of Electrical and Computer Engineering, Tech. Rep., October 2002.

[7] P. Y. O. William E. Green and G. Barrows, "Flying insect inspired vision for autonomous aerial robot maneuvers in near-earth environments," in *IEEE International Conference on Robotics and Automation (ICRA)*, vol. 3, New Orleans, LA, May 2004, pp. 2347–2352.

[8] T. McGee, R. Sengupta, and J. Hedrick, "Obstacle detection for small autonomous aircraft using sky segmentation," in *IEEE International Conference on Robotics and Automation*, 2005.

[9] S. E. Hrabar, P. I. Corke, G. S. Sukhatme, K. Usher, and J. M. Roberts, "Combined optic-flow and stereo-based navigation of urban canyons for a uav," in *Submitted to IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005.

[10] J. Zufferey and D. Floreano, "Toward 30-gram autonomous indoor aircraft: Vision-based obstacle avoidance and altitude control," in *IEEE International Conference on Robotics and Automation (ICRA'2005)*, 2005.

[11] J. G. G. Kenyon and R. Ziolkowski, "Time-to-collision estimation from motion based on primate visual processing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1279–1291, 2005.

[12] S. B. i Badia, "A fly-locust based neuronal control system applied to an unmanned aerial vehicle: the invertebrate neuronal principles for course stabilization, altitude control and collision avoidance," *The International Journal of Robotics Research*, vol. 26, no. 7, pp. 759–772, 2007.

[13] J. Oliensis, "A critique of structure-from-motion algorithms," *Computer Vision and Image Understanding (CVIU)*, vol. 80, no. 2, pp. 172–214, 2000.

[14] O. A. Takeo Kanade and Q. Ke, "Real-time and 3d vision for autonomous small and micro air vehicles," in *IEEE Conf. on Decision and Control (CDC 2004)*, December 2004, pp. 1655–1662.

[15] J. F. M. Srikanth Saripalli and G. Sukhatme, "Vision-based autonomous landing of an unmanned aerial vehicle," in *IEEE International Conference on Robotics and Automation*, 2002, pp. 2799–2804.

[16] O. Shakernia, R. Vidal, C. S. Sharpy, Y. Ma, and S. Sastry, "Multiple view motion estimation and control for landing an unmanned aerial vehicle," in *IEEE Conference on Robotics and Automation*, 2002.

[17] J. Santos-Victor and G. Sandini, "Uncalibrated obstacle detection using normal flow," *Machine Vision and Applications*, vol. 9, no. 3, pp. 130–137, 1996.

[18] R. B. Lorigo, L.M. and W. Grimson, "Visually-guided obstacle avoidance in unstructured environments," in *Proceedings of IROS '97*, Grenoble, France, September 1997, p. 373379.

[19] R. Nelson and Y. Aloimonos, "Obstacle avoidance using flow field divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 10, pp. 1102–1106, October 1989.

[20] T. Poggio, A. Verri, and V. Torre, "Green theorems and qualitative properties of the optical flow," MIT, Tech. Rep. 1289, 1991.

[21] N. Ancona and T. Poggio, "Optical flow from 1-d correlation: Application to a simple time-to-crash detector," *Int. J. Computer Vision*, vol. 14, no. 2, 1995.

[22] A. R. Z. Duric and J. Duncan, "The applicability of green's theorem to computation of rate of approach," *International Journal of Computer Vision*, vol. 31, no. 1, pp. 83–98, 1999.

[23] S. Maybank, "Apparent area of a rigid moving body," *Image and Vision Computing*, vol. 5, no. 2, pp. 111–113, 1987.

[24] T. Camus, "Calculating time-to-contact using real-time quantized optical flow," National Institute of Standards and Technology, Tech. Rep. NISTIR 5609, March 1995.

[25] T. Camus, D. Coombs, M. Herman, and T. Hong, "Real-time single-workstation obstacle avoidance using only wide-field flow divergence," *Videre: A Journal of Computer Vision Research*, vol. 1, no. 3, 1999.

[26] H. Liu, M. Herman, R. Chellappa, and T. Hong, "Image gradient evolution: A visual cue for collision avoidance," in *Proceedings of the International Conference on Pattern Recognition*, 1996.

[27] R. Cipolla and A. Blake, "Image divergence and deformation from closed curves," *International Journal of Robotics Research*, vol. 16, no. 1, pp. 77–96, 1997.

[28] A. Verri and T. Poggio, "Motion field and optical flow: Qualitative properties," *Pattern Analysis and Machine Intelligence*, vol. 11, no. 5, pp. 490–498, 1989.

[29] N. Hatsopoulos, F. Gabbiani, and G. Laurent, "Elementary computation of object approach by a wide-field visual neuron," *Science*, vol. 270, pp. 1000–1003, November 1995.

[30] F. Rind and D. Bramwell, "Neural network based on the input organization of an identified neuron signaling impending collision," *Journal of Neurophysiology*, vol. 75, p. 967 985, 1996.

[31] S. Yue and F. Rind, "A collision detection system for a mobile robot inspired by the locust visual system," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA05)*, April 2005, pp. 3832–3837.

[32] H. Okuno and T. Yagi, "Real-time robot vision for collision avoidance inspired by neuronal circuits of insects," in *IROS'07*, San Diego, CA, October 2007, pp. 1302–1307.

[33] L. Graham, "How not to get caught," *Nature Neuroscience*, vol. 5, pp. 1256 – 1257, 2002.

[34] S. Yue and F. Rind, "Collision detection in complex dynamic scenes using an lgmd-based visual neural network with feature enhancement," *IEEE transactions on neural networks*, vol. 17, no. 3, pp. 705–716, 2006.

[35] F. R. Shigang Yue, "Visual motion pattern extraction and fusion for collision detection in complex dynamic scenes," *Computer Vision and Image Understanding*, vol. 104, no. 1, pp. 48–60, October 2006.

[36] J. Craig, *Introduction to Robotics: Mechanics and Control*. Addison-Weslet Publishing Company, 1989.

[37] A. B. A. Criminisi, G. Cross and V. Kolmogorov, "Bilayer segmentation of live video," in *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2006.

[38] S. Birchfield and C. Tomasi, "Multiway cut for stereo and motion with slanted surfaces," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, September 1999.

[39] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient n-d image segmentation," *International Journal of Computer Vision (IJCV)*, vol. 70, no. 2, pp. 109–131, 2006.

[40] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in computer vision," in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, ser. LNCS, vol. 2134. Springer-Verlag, September 2001, pp. 359–374.

[41] V. K. C. Rother and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (SIGGRAPH'04)*, 2004.

[42] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 147–159, February 2004.

[43] R. Zabih and V. Kolmogorov, "Spatially coherent clustering with graph cuts," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR04)*, June 2004.

[44] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, November 2001.

[45] B. P. D. Greig and A. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society*, vol. 51, no. 2, pp. 271–279, 1989.

[46] Z. Wu and R. Leahy, "An optimal graph theoretic approach to data clustering: Theory and application to image segmentation," *IEEE trans. on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 1101–1113, November 1993.

[47] J. Lobo and J. Dias, "Relative pose calibration between visual and inertial sensors," *International Journal of Robotics Research*, 2007.

[48] J. Byrne and R. Mehra, "Wireless video noise classification for micro air vehicles," in *Association for Unmanned Vehicle Systems International (AUVSI'08)*, San Diego, CA, June 2008.

[49] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE trans. on Pattern Analysis and Machine Intelligence*, 1991.

[50] E. Simoncelli and W. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proceedings of the Second International Conference on Image Processing*, 1995.

[51] D. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information," *International Journal of Computer Vision*, vol. 5, no. 1, pp. 77–104, 1990.

[52] M. Lourakis and A. Argyros, "The design and implementation of a generic sparse bundle adjustment software package based on the levenberg-marquardt algorithm," Institute of Computer Science - FORTH, Heraklion, Crete, Greece, Tech. Rep. 340, Aug. 2004.

[53] B. Cohen and J. Byrne, "Inertial aided sift for visual collision estimation," in *IEEE International Conference on Robotics and Automation (ICRA'09)*, 2009.

[54] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1/2/3, pp. 7–42, April-June 2002.