2008-06-01

Application and Evaluation of Spatiotemporal Enhancement of Live Aerial Video using Temporally Local Mosaics

Dennis Eggett

Cameron Engh
cameron_engh@hotmail.com

Damon Gerhardt

Michael A. Goodrich
mike@cs.byu.edu

Bryan S. Morse

Original Publication Citation

B. S. Morse, D. Gerhardt, C. Engh, M. A. Goodrich, N. Rasmussen, D. Thornton, and D. Eggett. Application and Evaluation of Spatiotemporal Enhancement of Live Aerial Video using Temporally Local Mosaics. Proceedings of CVPR 28: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 28, Anchorage, Alaska.

BYU ScholarsArchive Citation

Application and Evaluation of Spatiotemporal Enhancement of Live Aerial Video using Temporally Local Mosaics

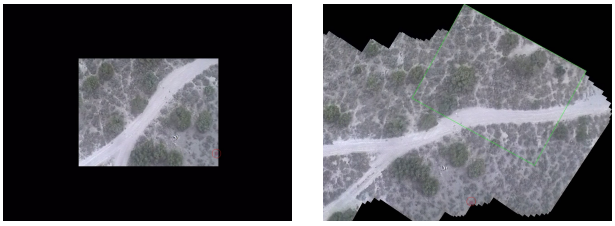
Bryan S. Morse, Damon Gerhardt, Cameron Engh, Michael A. Goodrich,
Nathan Rasmussen, Daniel Thornton, and Dennis Eggett
Brigham Young University, Provo, UT 84602
morse@byu.edu

Abstract

Camera-equipped mini-UAVs are popular for many applications, including search and surveillance, but video from them is commonly plagued with distracting jittery motions and disorienting rotations that make it difficult for human viewers to detect objects of interest and infer spatial relationships. For time-critical search situations there are also inherent tradeoffs between detection and search speed. These problems make the use of dynamic mosaics to expand the spatiotemporal properties of the video appealing. However, for many applications it may not be necessary to maintain full mosaics of all of the video but to mosaic and retain only a number of recent (temporally local) frames, still providing a larger field of view and effectively longer temporal view as well as natural stabilization and consistent orientation. This paper presents and evaluates a real-time system for displaying live video to human observers in search situations by using temporally local mosaics while avoiding masking effects from dropped or noisy frames. Its primary contribution is an empirical study of the effectiveness of using such methods for enhancing human detection of objects of interest, which shows that temporally local mosaics increase task performance and are easier for humans to use than non-mosaiced methods, including stabilized video.

1. Introduction

Using video feeds from inexpensive Mini Unmanned Air Vehicles (mUAVs) is becoming popular in a variety of applications, including search and rescue, military reconnaissance and target acquisition, counterterrorism, and border patrol. Their small size and ease of deployment enable mUAVs to gather up-to-date and high resolution surveillance that could be much more difficult to obtain otherwise. We focus here on the use of mUAV-acquired video for assisting search and rescue in remote wilderness areas (see [3] for a description of the field deployment of these mUAVs coordinated with human search teams), but the methods and study presented here apply to a variety of applications.



(a) Original video (b) Local mosaic video

Figure 1. Temporally local mosaics for enhanced detection. A small object in the non-mosaiced video (a), seen here at the lower-right corner of the frame and circled in red, may be visible in only a few frames. In the enhanced temporally local mosaic display (b) the same object, seen in the lower-middle of the view and again circled with red, is visible over nearly one hundred frames.

Unfortunately, video from mUAVs suffers from three general problems that make the video difficult for humans to interpret. First, the video tends to be shaky or jittery due to the small size and maneuverability of the aircraft, making it difficult for people to identify and track interesting features in the video. Second, it is easy to become disoriented while watching video from a mUAV, especially when the aircraft is circling or turning frequently, making it difficult to relate features seen throughout the video to each other or to real world features, locations, or objects. Third, in order to provide high enough resolution for detection and identification, UAVs will often use high-power lenses or fly lower to the ground, reducing the field of view and corresponding duration for which objects of interest are visible. This is further compounded when the UAV has to fly quickly due to the time-critical nature of search-and-rescue.

This paper proposes and evaluates the use of real-time temporally localized mosaics (Figure 1) to enhance the spatiotemporal display of live video for human observers in search applications. The enhanced spatial view increases situational awareness by giving observers a larger and more consistent sense of localization, and the enhanced temporal view increases an observer's opportunity to detect targets of interest. We present here and evaluate three alternative methods for video viewing based on temporally local mo-

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

saicing alone, stabilization and consistent orientation alone, or a hybrid stabilization/mosaicing approach.

The primary contribution of this paper is a quantitative human-observer study evaluating detection performance using these three display methods to each other and to unenhanced video while simultaneously performing a secondary task, similar to field use of these systems. We first describe the construction of these displays, then focus primarily on the evaluation of their use.

2. Related Work

Registering and combining multiple images together to form *mosaics*, larger images created by compositing multiple smaller source images, has been a well-studied problem in computer vision. This body of work is far too extensive to survey here, but we recommend [16] for an excellent recent survey and tutorial. Source images for mosaics can come from a variety of sources, including frames from a moving video camera to create larger views than obtainable through the field of view of any one frame (e.g., [10, 6]).

Video from UAVs can be used to create large-scale static mosaics of scenes from aerial imagery. (See [7] for an overview of this and related work.) [5] points out that this mosaicing of video simultaneously extends not only the spatial information but the temporal information (change capture) as well.

By updating the mosaic as new video is obtained, one can create a *dynamic mosaic* [6, 12] to display "live" content. The temporally local mosaics used here can be thought of as "memory-less" versions of these dynamic mosaics, in which we retain only the portion of the dynamic mosaic still in the display area.

Using combinations of telemetry, terrain models, reference imagery, and aerial video, one can further *georegister* the video to allow it to be overlaid in its larger context (terrain with reference imagery) [8, 19, 9], but this is beyond the scope of what we explore here, which is the effect such mosaicing has on detection of objects of interest. Obviously, once objects are detected and identified, such methods can help place them in proper geospatial context.

3. Methods

We first present three methods for enhancing video display for detection: local mosaicing, stabilization alone, and a hybrid stablization/mosaicing approach. For each method, we preprocess each frames on the fly, perform frame-to-frame alignment using video data only, then display them according to the viewing method used. Because the core methods build on well-established techniques, we focus primarily on application-specific constraints and on specific design and implementation details necessary for reproducibility of our empirical evaluation.

3.1. Video Acquisition

The mUAV platform is a custom airframe with a 50" wingspan. The sensor suite includes 3-axis rate gyroscopes, 3-axis accelerometers, static and differential barometric pressure sensors, a GPS module, and a 640×480 video camera on a gimballed mount. It uses a 900 MHz radio transceiver for data communication and an analog (NTSC) 2.4 GHz transmitter for video downlink. The UAV includes an onboard autopilot that provides stabilization of the aircrafts roll and pitch angles, attitude stabilization, altitude controller, and ability to fly to a waypoint. Because of the small size of the mUAV, no onboard processing of the video is performed.

3.2. Preprocessing

For all display methods, we preprocess each frame on the fly by deinterlacing it and correcting for radial distortion. Deinterlacing is essential for fast-moving mUAVs, as the translation between odd and even fields of the video is significant and introduces visible artifacts that impair detection as well as subsequent video processing. We have found it most efficient to simply use the 640×240 fields directly for the later vision algorithms, then upsample each to 640×480 frames for deinterlaced display.

3.3. Frame Alignment

To align successive frames of video, we use the Harris corner detector [4] to identify feature points, then establish correspondence between these points and use RANSAC [2] to estimate the Euclidean transformation (translation and rotation) between each pair of frames.

Although the small variation of the terrain within the camera's field of view relative to the mUAV's height above ground allows us to do as others have done and model the frame-to-frame transformation as a homography (e.g., [7]), our experience is that the inevitable accumulated error that occurs in all registration methods causes the scaling and skew components of the homography to adversely scale and distort the displayed frames over time. Because we are stabilizing video or creating temporally local mosaics in real time, not aligning entire sequences offline, we perform only frame-to-frame alignment and do not have the luxury of or a need for a global bundle adjustment [17].

Using only a Euclidean transformation can cause minor frame-to-frame alignment errors, but we have found these to be minimal compared to the need to display full-resolution, unskewed video for optimal detection by avoiding cumulative drift in scale or skewing. Since the drift in position and rotation (camera roll) is gradual and much longer than the time an object is visible in the video, we can still composite the frames and present them with a consistent orientation, even though the absolute orientation may drift

over time. This precludes geolocation or a true "north-up" display without the addition of heading telemetry from the mUAV, but this has no effect on the ability of an observer to detect objects of interest.

3.4. Display Methods

Once we have preprocessed and aligned the frames, we then display them to the user according to the display method selected.

3.4.1 Mosaiced Viewing

To allow the viewer a longer temporal window, we create a temporally local mosaic of recent video frames. This not only extends duration for which objects of interest are visible but also provides natural stabilization and maintains viewing orientation.

Let $\mathbf{Q}(t)$ denote the Euclidean transformation from frame $I(t-1)$ to the current frame $I(t)$. We can then define the cumulative transformation $\mathbf{Q}'(t)$ from frame $I(0)$ to frame $I(t)$ as the composition of these transformations using the following recurrence relation:

$$\mathbf{Q}'(0) = \mathbf{I} \quad (1)$$

$$\mathbf{Q}'(t) = \mathbf{Q}(t)\,\mathbf{Q}'(t-1) \quad (2)$$

For our temporally local mosaic, we build a composite image I' that is the size of the viewing window and keep in it only the frames that are still visible in the current display. We begin each video display by copying the first frame $I(0)$ onto the center of $I'(0)$ and displaying it to the user. We then align each successive frame $I(t)$ to its predecessor $I(t-1)$, warp frame $I(t)$ using the cumulative transformation $\mathbf{Q}'(t)$ to the corresponding position and orientation, and superimpose this frame onto the existing $I'(t-1)$ to create $I'(t)$. Because of the temporal ordering of the frames and the desire to always show in full the current frame, no frame-to-frame blending is used.

When the warped position of the new frame $I(t)$ exceeds the bounds of the viewing window, we translate $I'(t-1)$ by the minimum amount necessary so that the new frame can be added in full. Denoting this required offset (if any) as $(x(t), y(t))$ and the corresponding translation matrix as $\mathbf{T}(t)$, we can then denote the warp $\mathbf{M}(t)$ used to apply each frame to the already-translated mosaic $I'(t-1)$ as

$$\mathbf{M}(t) = \mathbf{Q}'(t)\,\mathbf{T}(t) \quad (3)$$

Intuitively, the $\mathbf{Q}'(t)$ component of $\mathbf{M}(t)$ tracks the motion of the video while $\mathbf{T}(t)$ accounts for the required scrolling of the display so that each frame is placed correctly.

For forward-moving mUAVs with downward pointing cameras, this causes the display to pan steadily with the direction of motion (Figure 2(a)). As past frames composited onto I' move outside of the viewing display, their content is no longer retained. This simplifies the process of compositing the video while still providing the desired temporal expansion. It also avoids awkward accumulated-error artifacts when looping back over previously-seen areas or the need to test for and compensate for such artifacts (e.g, [15]).

3.4.2 Stabilized Viewing with Consistent Orientation

To see whether the effects of temporally local mosaics for display are due to the mosaicing or solely to the stabilization and consistently oriented display, we compare them to stabilized, consistently-oriented video without mosaicing.

To stabilize the display of the image, we calculate the warped positions of the center of each image and fit a spline to the warped positions. This dampens the jittering caused by erratic motions of the small UAV while allowing general progression of the plane's search/flight path.

Let $\mathbf{c}(0)$ denote the position of the center of the image. We can then build the paths of the warped centers $\mathbf{c}(t)$ of the images as

$$\mathbf{c}(t) = \mathbf{Q}'(t)\,\mathbf{c}_0 \quad (4)$$

and use the most recent n transformed centers as controls points $\mathbf{b}_i$ for a spline B:

$$\mathbf{b}_i = \mathbf{c}(t-n+i),\ 0 \le i \le n \quad (5)$$

We define the smoothed location $\mathbf{q}(t)$ as the midpoint of the spline between $\mathbf{c}(t)$ and $\mathbf{c}(t-n)$, which we evaluate using de Casteljau's mid-point algorithm [1].

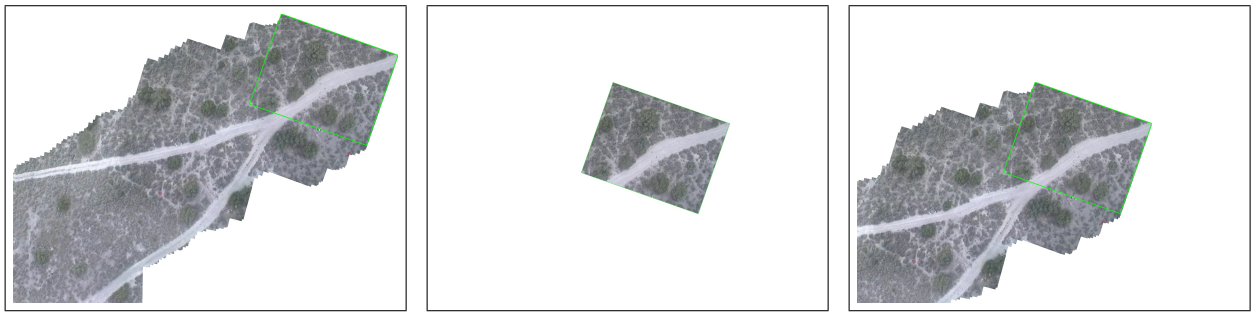
We then warp the image to the displayed position using the transformation

$$\mathbf{A}(t) = \mathbf{Q}'(t)\,\mathbf{T}(t)\,\mathbf{S}(t) \quad (6)$$

where $\mathbf{S}(t)$ causes translation by $\mathbf{c}(t)-\mathbf{q}(t)$ and $\mathbf{T}(t)$ is calculated as for the local mosaic.Thus, the stablilizing transformation $\mathbf{A}(t)$ for frame t is a composite of the cumulative history $\mathbf{Q}'(t)$ tracking the moving center $\mathbf{c}'(t)$ of the image, the translations $\mathbf{T}(t)$ that would be used to keep this frame within view, and a stabilizing shift $\mathbf{c}(t)-\mathbf{q}(t)$. The relationships between these components are illustrated in Figure 3.

Since the stabilized view uses $\mathbf{Q}'(t)$ and translations only, the orientation of the currently displayed frame is the rotation component of $\mathbf{Q}'(t)$. This provides the same consistent orientation as with the temporally local mosaic.

The effect of this stabilized display is that the content of the video moves smoothly and in a consistent orientation between frames, while the outer bounding box of the displayed frame may jitter and rotate within a larger display (Figure 2(b)). While it would be possible to crop the frame so that the jittering of the bounds of the displayed frame would not be visible, as is commonly done with stabilized



(a) Local Mosaic (b) Stabilized Only (c) Stabilized Mosaic

Figure 2. View presentation methods. Example of the stabilized mosaic and stable mosaic view presentations, each with a view three times the size of the original capture frame size.

displays, we avoid this approach because we want to provide to the observer the full area seen by the current frame and not lose video content. We also avoid using methods for full-frame stabilization [14, 18, 11], which synthesize data from the current and surrounding frames to provide a full field of view, because we want to display only actual rather than synthesized content.

3.4.3 Stabilized/Mosaic Viewing

One effect of the temporally local mosaic translating the display only when the current frame exceeds the viewing window is that when the video jitters it may cause the display to pan, then it may move back into the displayable area, then it may shift so as to cause the display to pan again. This causes jerky motion of the entire display as the current frame unpredictably "knocks" against the edge of the displayed area. To compensate for this, we combine the local mosaic and the stabilized presentation to create a hybrid stabilized mosaic (Fig 2(c)) by warping each frame as with the stabilized view ($\mathbf{A}(t)$), then compositing the new frame onto the translated $I'(t-1)$ rather than clearing previous frames.

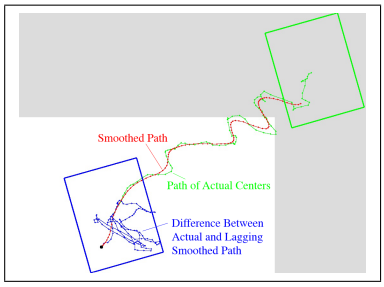
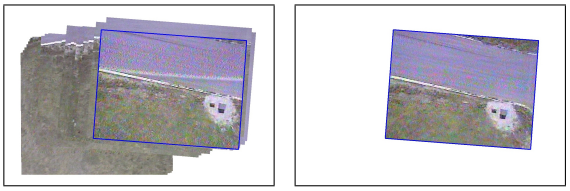


Figure 3. Stabilization path. As the tracked center of the current image moves (green), a spline is fit to the last n tracked frame centers (red). The difference between the lagging smoothed position and the actual position of the center of the image, as well as additional translations used to keep the entire display within view, are used to create the stabilized path of the displayed frames (blue).



(a) Invalid registration indication (b) Mosaic reset

Figure 4. Indication of data loss. When frames are corrupted, we drop them and change the highlight of the current frame to blue (a) to indicate the data loss. After a predefined limit of lost frames, we reset the display and begin building the mosaic again (b).

3.5. Interface Issues

We have found it useful to highlight the boundary of the current frame so as to distinguish for the user the "live" frame and the recent history (Figure 2).

When frames are corrupted due to transmission problems, as are inevitable when transmitting analog video from mUAVs, we may have frames that cannot be accurately aligned to the previous frame. We detect these corrupted frames by testing the accuracy of the alignment (number of corresponding points in the RANSAC consensus set) and simply do not display these frames, continuing to display the last well-aligned frame and aligning future frames to it. Not only does this create a more pleasing display due to the removal of noisy frames, but it specifically avoids well-known visual masking effects in human perception that can interfere with detection of briefly seen objects [13] .

While not displaying the corrupted frame(s), we change the color of the boundary highlight for the last displayed frame so as to indicate the data loss (Figure 4). Should the number of lost frames exceed a predefined limit (we use five frames), we clear the display and restart the presentation of frames by treating the next frame as $I(0)$.

4. User Study Design

We here present the design for a user study we conducted to quantify the effectiveness of these different presentation methods at enhancing subjects' abilities to detect and iden-

tify objects of interest within mUAV-acquired video while performing a secondary visual task. These tasks were designed to mimic common tasks performed in scenarios in which these mUAVs have been used. In addition to the three forms of enhanced video presented in Section 3, we also include the de-interlaced and undistorted but otherwise unenhanced original video.

The study involved 14 naïve and 12 potentially biased volunteer subjects. Each participant was asked to perform two tasks simultaneously in a controlled scenario over 16 different trials. Prior to completing the trials, each subject was required to complete training, during which they were introduced to the tasks and allowed to practice them before proceeding to the measured trials.

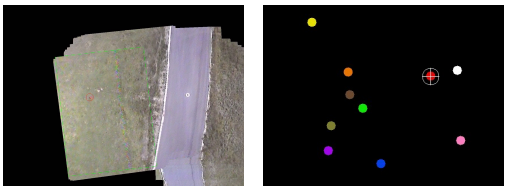
On the primary video display (Figure 5(a)), each subject was presented with a controlled random ordering of 16 different short video clips acquired using a mUAV engaged in common search patterns. Each clip lasted about 1.5 minutes and was presented to the subject using one of the four possible views: original, stabilized, mosaic, or stable mosaic. Within each clip, there were a random number of objects of interest (faint red umbrellas) placed randomly in the scene, which the subjects were asked to detect and identify.

On the secondary task display (Figure 5(b)), subjects were shown a controlled random set of uniquely colored spots dependant on the corresponding video clip. We regenerated a new display with 10 spots every 2–5 seconds, using 12 unique colors. Of these, subjects were asked to detect and identify as many red spots as possible without jeopardizing their ability to detect and identify objects of interest in the primary video display. This task was designed to provide a measure of the subject's ability to simultaneously perform a task similar to that traditionally required for operating the aircraft while performing a video search.

In order to facilitate within- and between-subject video clip and display method (view) comparisons, we use a counterbalanced design ensuring that every clip and every viewing method is seen an equal number of times per subject as well as seen a progressively equal number of times by each subject. Across all subjects, each view was seen a total of 104 times and contained a total of 254 red umbrellas.

Each clip-view combination was presented using an interface that allowed the subjects to easily select objects of interest in the video and secondary displays. This was designed to require as little training as possible so as to minimize performance differences between naïve and biased subjects. The subjects interacted with the system using only the mouse, and the controls for the video display and the secondary display were intentionally similar.

In order to decouple subjects' hand-eye coordination from the detection task, any time an object of interest was thought to be seen on the primary video display, they could freeze the frame by mouse-left-clicking anywhere in the display window. Freezing the frame caused the display of the video to freeze but did not pause the video playback, which continued in the background. The longer the frame was frozen, the more video content the user would miss seeing, just as in a live search situation. From there, the user could either left-click on the object of interest to identify it or right-click to indicate no selection, with either action resuming playback. The secondary task display used a similar control interface, allowing subjects to place, adjust, or clear markers indicating a detected red spot.



(a) Video Display (b) Secondary Display

Figure 5. User study displays presented using dual monitors.

After each trial, each subject was asked to answer three post-trial questions presented on the secondary display. These questions related their perception of their performance on the trial and to their relative preference between the just-finished trial and the previous one. After completing all 16 trials, each subject was asked to complete a brief questionnaire about their overall impressions and preferences.

5. Results and Discussion

In the described user study, we gathered hit rates for the primary and secondary tasks, hit rates given whether the subject is biased or naïve, hit rates within the current frame, hit rates within the history of the mosaic (when applicable), and false-positive rates and types. We also gathered subjective feedback from the participants.

Three preliminary observations are in order before comparing the main results. First, there is no statistical difference between the objective results for naïve and biased participants, who had hit rates for the primary task of 73% and 72% respectively (Table 1). Thus, the remaining analysis does not distinguish between the two participant types. Second, detection and identification success rates for the secondary display are very high and consistent across all participants and all views at about 94% (Table 2). This suggests that any influence from the additional cognitive load on the results will be expressed mainly in the differences among detection rates within the primary video display. Third, one particular video clip was an outlier wherein all participants identified all of the objects of interest regardless of the accompanying view. This clip was removed from the analysis.

	ω	P	% improvement
mosaic	1.6610	84.04%	45.33%
stable mosaic	1.5486	82.47%	42.62%
stabilized	0.3935	59.71%	3.26%
unenhanced	0.3156	57.83%	0.00%
biased	1.0051	73.21%	-
naïve	0.9543	72.20%	-

Table 1. Hit probability comparisons among the different presentation views as well as between the naïve and biased subjects, where ω is the least-squares means estimate and P is $(e^{\omega})/(1+e^{\omega})$, *i.e.*, the probability that the object of interest will be detected given the corresponding presentation view or subject. The improvement over P_{low}, the unenhanced view, was computed by $(P_{view} - P_{low})/P_{low}$.

	Spot Hit Rate
mosaic	94.88%
stable mosaic	93.24%
stabilized	93.67%
unenhanced	94.99%

Table 2. Performance at the secondary task by presentation view.

5.1. Primary Task Performance

The detection hit rates for the primary task are shown in Table 1 and support our hypothesis that extending the spatiotemporal view using local mosaics increases the probability that objects of interest will be detected throughout mUAV-acquired video. The mosaic view gave the largest increased percentage at 45.33% in hit probability over the unenhanced view. Also, there is a strong $(\sim 43\%)$ improvement from the non-mosaiced to mosaiced views.

Comparisons of similarity obtained by a least-squares means indicate that the two mosaiced views were similar to each other ($p = 0.9674$) and that the two non-mosaiced views were likewise similar ($p = 0.9804$), but that the mosaiced and non-mosaiced forms were statistically different ($p < 0.00001$). These results were obtained via a multiple-comparison ANOVA with the Tukey-Kramer adjustment.

This improvement is largely explained by examples such as that in Figure 1. In the original video (1(a)), the object of interest (red umbrella) is visible for only a few frames in the lower-right corner of the original view, and would appear similarly in the stablized view. However, in the corresponding mosaicked view (1(b)), the object is visible for a much longer time over possibly hundreds of frames, or several seconds, before it moves out of the viewing frustum.

5.2. Hits in the Current Frame vs. Mosaiced History

To see whether subjects were actually detecting the objects in the current frame or in the mosaiced history (if applicable), we also recorded hits in each area respectively, the results of which may be seen in Table 3.

It is interesting to note the connection between the increases in hit probabilities between the mosaiced and non-mosaiced presentation views shown in both Table 1 and the "History" column of Table 3. We believe that this reinforces

	Current Frame		History		Total
mosaic	128	62.44%	77	37.56%	205
stable mosaic	137	68.66%	62	31.34%	199
stabilized	147	100.00%		n/a	147
unenhanced	144	100.00%		n/a	144

Table 3. Hits in the current frame vs. hits in the mosaiced history.

that the increase in hit probability is largely due to the provision of a history in the mosaiced presentation views.

We also believe that the main difference in hit probabilities between the stable mosaic and mosaic views is because the stable mosaic view presents less of a history than the mosaic view.

5.3. Analysis of Types of Misses

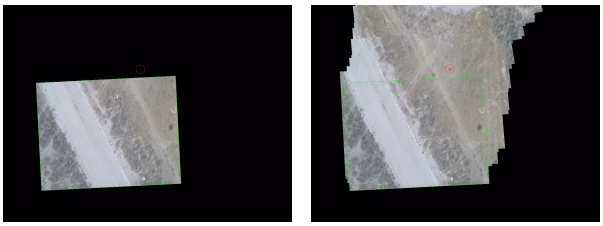
To further determine whether the mosaic's history was providing the increased detection, we examined the types of misses that occurred. Did the subject completely miss the object (not detected) or did they see it and click on it only after it left the current frame (late hit)? (See Figure 6.) By separating out the late hits from those targets that were not detected (Table 4), we see a significant difference in the types of misses that occurred when using the two mosaiced views compared to the two non-mosaiced views. The mosaic view, which displays the longest history, had the fewest number of late hits; the stabilized mosaic, which has a shorter history, had more late hits; and the two non-mosaiced views had more than four times the number of late hits. The greater number targets not detected in the non-mosaiced views indicates that these targets were completely missed.

5.4. False Positives

Table 5 shows an increase in the number of false positives (FPs) that occurred using the mosaiced views. FPs

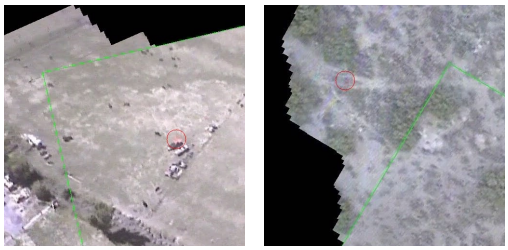
	Late Hits		Not Detected		Total
mosaic	2	4.08%	47	95.92%	49
stable mosaic	7	13.21%	46	86.79%	53
stabilized	33	31.13%	73	68.87%	106
unenhanced	31	27.93%	80	72.07%	111

Table 4. Classification of the misses by presentation view.



(a) A stabilized view's "Late Hit" (b) A stable mosaic view's hit of the same frame

Figure 6. Example of a late hit. A user may detect but not react quickly enough with a non-mosaiced view (a) but may be able to accurately detect and localize an object in a mosaiced history (b).



(a) In the current frame (b) In the mosaic

Figure 7. Examples of believable false positives. All methods are equally likely to generate FPs for confounding objects in the current frame (a). Once they pass from the current frame, they are more likely to be detected as FPs in the mosaiced views (b).

	FP total		Current		Late / History	
mosaic	19	18.27%	7	6.73%	12	11.54%
stable mosaic	11	10.58%	7	6.73%	4	3.85%
stabilized	6	5.77%	4	3.85%	2	1.92%
unenhanced	9	8.65%	7	6.73%	2	1.92%

Table 5. False positives in current frame and in mosaiced history.

can occur to the fault of a mosaic presentation mainly due to possible noise caused by the video transmission or capture device, or due to possible misalignments in the mosaic. These would be manifest as FPs made in the history; and according to Table 5, our results show a significant increase in FP's in the history of the mosaic view over those made in the surrounding area of the unenhanced view. They also show a 4% chance of having a FP occurrence in the history given the stable mosaic view, and a 12% chance given the more lengthy history presented in the mosaic view.

Reexamination of the video segments in which the FPs occurred showed that there were three kinds: (1) those occurring in the current frame (Figure 7(a)), which are likely to occur regardless of the view presentation; (2) those occurring in the history due to an alignment error, which can be attributed to the mosaicing; and (3) true objects or transmission noise bursts in the scene that appear similar to the intended objects of interest (Figure 7(b)), which are more likely to generate FPs in the history of the mosaiced views, not as a mosaicing artifact but due to demonstrated increased detectability in these views. In this last case, mosaiced views cause more false positives just as they cause more hits because potential confounding objects are themselves now more detectable.

5.5. Subjective Results

After each trial beyond the first one, each subject was asked to compare the presentation view of that trial to the one preceding it and rank it as harder ($<$), about the same ($\sim$), or easier ($>$). The collective results are shown in Table 6. We found no statistical bias in preference due to ordering, so we present here only the combined results for all times where one method was preferred to another, regardless of whether it preceded or followed the other. The

	mosaic	stable mosaic	stabilized	unenhanced	mosaic	stable mosaic	stabilized	unenhanced
mosaic	3	26	7	16	2	26	**53**	**41**
stable mosaic	-	6	9	13	8	2	**44**	**32**
stabilized	-	-	5	23	1	14	8	10
unenhanced	-	-	-	7	5	11	15	2
	Row $\sim$ Column				Row $>$ Column			

Table 6. Subjective pairwise comparison of the presentation views. Values indicate how many times users found one method similar to another ($\sim$) or preferred one to the other ($>$).

	Umbrellas			Spots		
	-	0	+	-	0	+
mosaic	10	64	30	21	49	34
stable mosaic	14	61	29	23	44	37
stabilized	11	44	40	20	44	40
unenhanced	9	44	49	19	53	32

Table 7. Hit confidence measures. Values indicate the number of times the subjects underestimated ("-"), accurately estimated (0), or overestimated the number of misses ("+").

results show an obvious heavy leaning towards the easiness of the mosaiced views over the non-mosaiced views.

One unexpected result is that the stabilized view seems to have been perceived as more difficult than the unenhanced view, and the stable mosaic view similarly more difficult than the mosaic view. Participant comments suggested that this difficulty was heavily influenced by the visual secondary task that required the subject to be visually engaged on another screen. When they were forced to look away from the video display for a moment and then look back again, the current frame was in a much less predictable position on the video display. Because of the moving high-gradient boundary of the displayed frames, users' attention was drawn to the moving frame boundary, especially when looking back to the display but even to some degree when not shifting between displays.

5.6. Subject Confidence

As a final subjective measure, we asked the subjects to estimate their own effectiveness in the tasks performed by asking them to report the number of spots and umbrellas they thought they missed during each trial. In a way similar to analyzing the late-hit misses, this allows us to see whether the subjects simply missed targets or saw them but didn't indicate them correctly. We then counted how many times the subjects underestimated ("-"), accurately estimated ("0"), or overestimated ("+") the number of misses (Table 7). The results suggested that the subjects tended to consistently underestimate the number of misses, *i.e.*, when they missed a target, they were usually not aware of it, but this was reduced when using mosaiced views.

6. Conclusion

Enhancing mUAV-acquired video for human observers in real-time search applications imposes a number of constraints on the processing of the video: real-time vision algorithms, high detection rates, handling of noisy or lost frames, consistency of relative orientation, and consideration of human perception and interface factors.

The empirical evaluation presented here shows that the use of temporally local mosaics can significantly increase the detectability of objects of interest and ease of use for human observers. The two mosaiced display methods presented here (mosaicing alone or combined with stabilization) were nearly identical in their impact on performance and significantly better than either of the two non-mosaiced display methods (unenhanced or stabilized alone).

Mosaiced displays also bring a slight increase in the number of false positives, some of which are due to artifacts in the mosaicing, but many of which are simply due to the fact that potential confounding objects are themselves likewise made more detectable.

Study participants were also more likely to be overconfident when using the non-mosaiced forms of display than when using the mosaiced forms. This has significant implications in search situations because it suggests that once an area is searched we may be too mistakenly confident that the area covered does not contain an object of interest.

Stabilizing the non-mosaiced video surprisingly seems not to improve detection, nor did stabilizing the current frame in a mosaiced presentation. Study participants suggested that these forms of non-cropped stabilized display were actually harder to use, perhaps due in large part to the presence of a secondary visual task that required them to continually shift their attention between the video and another display. Not cropping the stabilized video allows the display to maintain full information but introduces a high-gradient moving boundary, which seems to make the task more difficult, especially when shifting attention.

The study presented here did not assess the influence of non-visual secondary tasks, nor did it attempt to assess the impact of these display methods on long-term fatigue. Each of these would be interesting studies to pursue in the future.

Although studied here in the context of search-and-rescue applications, these findings should extend to video presentation for other applications of mUAVs and similar video sources.

Acknowledgments

The work was partially supported by the National Science Foundation under grant number 0534736. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] P. de Casteljau. Courbes et surfaces à pôles. Technical report, Citroën, Paris, 1959.

[2] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[3] M. A. Goodrich, B. S. Morse, D. Gerhardt, J. L. Cooper, J. Adams, C. Humphrey, and M. Quigley. Supporting wilderness search and rescue using a camera-equipped mini UAV. *Journal of Field Robotics*, 25(1):89–110, January 2008.

[4] C. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Converence, Manchester, UK*, pages 147–151, August 1988.

[5] M. Irani and P. Anandan. Video indexing based on mosaic representations. *Proceedings of the IEEE*, 86(5):905–921, 1998.

[6] M. Irani, P. Anandan, and S. Hsu. Mosaic based representations of video sequences and their applications. In *ICCV '95: Proceedings of the Fifth International Conference on Computer Vision*, pages 605–611, 1995.

[7] R. Kumar, H. Sawhney, S. Samarasekera, S. Hsu, H. Tao, Y. Guo, K. Hanna, A. Pose, R. Wildes, D. Hirvonen, M. Hansen, and P. Burt. Aerial video surveillance and exploitation. *Proceedings of the IEEE: Special Issue on Third Generation Surveillance Systems*, 89(10):1518–1539, October 2001.

[8] R. Kumar, H. S. Sawhney, J. C. Asmuth, A. Pope, and S. Hsu. Registration of video to geo-referenced imagery. *Proceedings of IEEE CVPR*, pages 54–62, August 1998.

[9] Y. Lin and G. Medioni. Map-enhanced UAV image sequence registration and synchronization of multiple image sequences. In *CVPR '07: IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2007.

[10] S. Mann and R. W. Picard. Virtual bellows: constructing high quality stills from video. In *IEEE International Conference on Image Processing*, volume 1, pages 363–367, 1994.

[11] Y. Matsushita, E. Ofek, X. Tang, and H.-Y. Shum. Full-frame video stabilization. In *CVPR (1)*, pages 50–57. IEEE Computer Society, 2005.

[12] H. Nicolas. New methods for dynamic mosaicking. *Image Processing, IEEE Transactions on*, 10(8):1239–1251, 2001.

[13] H. Piéron. *The Sensations: Their Functions, Processes, and Mechanisms*. Yale University Press, 1952.

[14] A. Rav-Acha, Y. Pritch, D. Lischinski, and S. Peleg. Dynamosaicing: Mosaicing of dynamic scenes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1789–1801, 2007.

[15] H. S. Sawhney, S. Hsu, and R. Kumar. Robust video mosaicing through topology inference and local to global alignment. In *ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume II*, pages 103–119, London, UK, 1998. Springer-Verlag.

[16] R. Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 2(1):1–109, 2006.

[17] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment—a modern synthesis. In B. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, volume 1883 of *Lecture Notes in Computer Science*, pages 298–372. Springer-Verlag, 2000.

[18] Y. Wexler, E. Shechtman, and M. Irani. Space-time completion of video. *Transactions on Pattern Analysis and Machine Intelligence*, 29(3):463–476, 2007.

[19] R. P. Wildes, D. J. Hirvonen, S. C. Hsu, R. Kumar, W. B. Lehman, B. Matei, and W. Y. Zhao. Video georegistration: algorithm and quantitative evaluation. In *ICCV '01: Proceedings of the Eighth IEEE International Conference on Computer Vision*, volume 2, pages 343–350 vol.2, 2001.