Fusion-Based Approach for Long-Range Night-Time Facial Recognition

Robert B. Martin, Mikhail Sluch, Kristopher M. Kafka, Andrew Dolby, Robert Ice and Brian E. Lemoff
WVHTC Foundation, 1000 Technology Drive, Suite 1000, Fairmont, WV, USA 26554

ABSTRACT

Long range identification using facial recognition is being pursued as a valuable surveillance tool. The capability to perform this task covertly and in total darkness greatly enhances the operators’ ability to maintain a large distance between themselves and a possible hostile target. An active-SWIR video imaging system has been developed to produce high-quality long-range night/day facial imagery for this purpose. Most facial recognition techniques match a single input probe image against a gallery of possible match candidates. When resolution, wavelength, and uncontrolled conditions reduce the accuracy of single-image matching, multiple probe images of the same subject can be matched to the watch-list and the results fused to increase accuracy. If multiple probe images are acquired from video over a short period of time, the high correlation between the images tends to produce similar matching results, which should reduce the benefit of the fusion. In contrast, fusing matching results from multiple images acquired over a longer period of time, where the images show more variability, should produce a more accurate result. In general, image variables could include pose angle, field-of-view, lighting condition, facial expression, target to sensor distance, contrast, and image background. Long-range short wave infrared (SWIR) video was used to generate probe image datasets containing different levels of variability. Face matching results for each image in each dataset were fused, and the results compared.

Keywords: Image Fusion, Face Recognition, SWIR, Night Vision, Surveillance, Biometrics, Active Imaging

1. INTRODUCTION

As tensions increase in areas around the world, new technologies are continually being created and evaluated to protect oneself or property. Video surveillance systems are very popular because they put a standoff distance between the target and the operator, can be automated for 24 hour operation, and also can be used for post-processing. When teamed with add-on software packages, these surveillance systems become a very formidable tool for defense applications. Facial recognition systems are an example of this and are becoming increasingly popular in many communities including the military, commercial, and private sectors. As the desire for this technology grows, so does the requirement for its accuracy. The ability to detect and identify individuals at long ranges, with high accuracy, in short periods of time, whether day or night, in all weather conditions, while remaining completely tactical and covert would prove highly favorable in this application. For this reason, the West Virginia High Technology Consortium Foundation (WVHTCF), under a research contract from the Office of Naval Research (ONR), with funding and oversight from the Office of the Secretary of Defense Deployable Force Protection Program (DFP), is developing the Tactical Imager for Night/Day Extended Range Surveillance (TINDERS)\(^1,2,3\). This is an actively illuminated short-wave infrared (SWIR) video imaging system that uses illumination that it invisible to conventional silicon optics and the naked eye. This system provides imagery suitable for facial recognition at ranges of up to 400 meters in complete darkness and imagery for tracking purposes beyond three kilometers. Optimization of the face recognition performance can be pursued in three primary ways. First, the optical hardware can be optimized to produce the highest quality SWIR imagery. Second, the face matching algorithms can be optimized to produce the most accurate matching of a single SWIR probe image to a database of visible-spectrum mug-shots. Most of the effort to date has focused on these first two approaches. The third approach, which is discussed in this paper, is to fuse the matching results for multiple SWIR probe images of the same person to increase the identification accuracy. In principle, the longer an individual is monitored by the video surveillance system, the more facial images can be captured and processed, and the more accurate the fused matching result will be. This paper investigates the hypothesis that it is more beneficial to have a low correlation than a high correlation between the fused probe images. The data suggests that the fusion of multiple images with low correlation results in a somewhat larger improvement in accuracy than the fusion of multiple images with high correlation. The results of this work may ultimately lead to the integration of new features in the TINDERS system, and may also be generalizable to other face recognition systems based on video surveillance.
2. SYSTEM DESCRIPTION

2.1 System Overview

In 2009, TINDERS began as a laboratory prototype providing for initial proof of concept. After successful laboratory demonstration, this unit was mobilized and transported across the country where it demonstrated long range person detection and identification in complete darkness. The success of this system was then improved upon with the development of the second generation system. Improvements were made in size, weight, durability, mobility, usability, and ruggedness.

TINDERS functions as a night/day video surveillance system that can be deployed as a standalone system or as part of a larger network. It can be tasked or “slewed to” a specific location of interest. The system provides live video output allowing for either manual or automatic person detection to be accomplished. Once a person is detected, positive identification can be attempted. The TINDERS system has a large storage capacity allowing for the post-processing of recorded SWIR video data.

2.2 SWIR Illumination

TINDERS is an actively illuminated SWIR system with optical zoom capability. The illumination beam divergence angle is matched to the imager’s field-of-view (FOV) with motorized optics. The source is a fiber coupled superluminescent light emitting diode (SLED) fed to an erbium-doped fiber amplifier, running at constant power. Operating within a common telecommunication wavelength band that is greater than 1400 nm, the TINDERS system is completely invisible to the naked eye and conventional silicon imagers. The system remains completely eye safe at all times as defined by the ANSI Z136 and IEC 60825 laser eye safety standards\(^4\). The eye-safety classification is Class 1M at the illuminator output and Class 1 at the target. Safe operation at this wavelength allows for 65 times more light when compared to 800 nm.

While typical visible spectrum facial recognition has made major advances, operation in the SWIR band does offer some challenges. As illustrated in Figure 1, noticeable differences between SWIR and visible-spectrum imagery include low skin reflectivity, lack of a noticeable pupil, and high hair reflectivity. However, the landmarks, such as the edges of the nose, lips, and eyes remain the same as in the visible spectrum. For a system to be most useful, it must match its acquired sensor data against existing databases. For TINDERS, the challenge is to match the SWIR facial imagery against an existing database of visible spectrum mug shots.

Figure 1. Visible Spectrum and SWIR Mug Shot . (left) Visible spectrum mug shot (right) SWIR probe image.

2.3 Hardware

The TINDERS system is comprised of three components: the optical head unit, the electronics enclosure, and the control computer. Figure 2 depicts both a conceptual illustration of the components and a picture of the deployable unit.
The optical head unit contains all the optics including the mechanisms to move them. Included within is a laser range finder, a focal plane array, and a video processing board. All components are encased in a climate controlled and environmentally sealed enclosure mounted to a FLIR® pan/tilt unit. Typically it is deployed on the tripod as depicted above but can also be mounted on a mast or tower.

The electronics enclosure contains the supporting power supplies, communication electronics, and illumination source. As previously mentioned, this is a fiber-coupled system, meaning the SLED, amplifier, and filter are contained within this thermal electrically cooled enclosure. The Mil Spec connections facilitate 24 hour operation in all conditions.

The system is currently controlled by a semi-rugged computer. The computer does not need to be co-located with the system. Both simply need to be plugged into any ordinary 100 Mbit/s network and 110 VAC. The unit offers a large storage capacity for video post-processing as well as in situ operation.

2.4 Control Software

TINDERS can be controlled by an operator from the control computer console or a client running Windows Remote Desktop Connection if the system is network connected. The operator is presented with a number of controls to perform their desired task whether it is surveillance, detection, identification, or tracking. The system has integrated person and face detection algorithms allowing for tracks to be initiated with little operator involvement. Once these are initiated, the operator then can attempt facial recognition as discussed in the next section. TINDERS, as previously mentioned, can also be integrated into a larger network of sensors and report its findings to a database or central location for further scrutiny.

2.5 Facial Recognition

The TINDERS control software and the facial recognition package are both contained within one graphical user interface (GUI). The facial recognition platform is powered by a commercially available package developed by MorphoTrust USA. It is based on their Face Examiner Workstation5, but with the addition of custom software that pre-processes the SWIR images before they are matched against a gallery that is pre-populated with a database of visible-spectrum mug shots.

Selection of video frames for face recognition can be done in either an automated or manual mode. In the automated mode, faces are detected in video and automatically submitted to the face recognition process. In manual mode, the operator clicks a button to cue a predefined number of SWIR video frames to be submitted to the facial recognition process. Alternatively, the user can load previously-saved files containing SWIR facial imagery. The user can then manually mark the eye positions. Matching is then initiated. Each probe image is then matched against each gallery image and assigned a score. These are then ranked based on the fused score.
Figure 3 illustrates the TINDERS GUI with integrated facial recognition when used in manual mode. The image on the left represents the live video feed. The image on the right is a probe image that was searched against the visible spectrum database. The visible images below show the possible matches with the correct subject scoring Rank 1.

![Screenshot of the TINDERS GUI, showing user controls, facial recognition in “manual mode”, correct match, and other controls.](image)

2.6  FUSION

The fusion process involves combining the face matching results from multiple probe images to increase the identification confidence level. The live video feed supplied by the TINDERS system supplies up to 30 frames per second of SWIR facial imagery. As long as the images belong to the same person, they can continue to be matched and the matching results fused. For the results presented in this paper, the “maximum score” fusion technique is used in which each visible gallery image is assigned a fused matching score equal to its highest single-probe matching score across all of the probe images.

The probe images submitted to facial recognition do not necessarily have to be identical to one another; however, in most TINDERS experiments that have been analyzed to date, test subjects were stationary and facing the camera with a neutral facial expression, having a pose angle and facial expression very similar to the enrolled visible-spectrum gallery images. As a result, the fusion process in these experiments has combined the matching results for a set of probe images that closely resemble one another, i.e. probe images with high correlation. While the fused results have always been more accurate than the single-probe results, it is possible that the lack of variability in the probe images has limited the performance improvement provided by fusion. Thus, an experiment was performed to investigate the impact of probe image variability on fusion performance.

3.  EXPERIMENTAL DESCRIPTION

3.1  Database

The visible-spectrum gallery used for this experiment included enrollment images from two datasets. The first dataset was collected in summer through winter of 2012 in conjunction with a research group at West Virginia University (WVU). This dataset consists of 104 individuals. The second dataset, which was collected specifically for this experiment, included 10 individuals. Thus, the total database size used in this experiment was 114 faces. One individual inadvertently appeared in both datasets; however, this does not affect our ability to test the hypothesis, as the duplicate record affects all matching scores in the same way.

3.2  Experimental Set Up

In order to determine whether it is more favorable to have a fusion of images of high or low correlation, sets of images need to be produced having both high and low correlation. These sets need to be produced in a repeatable manner. A
test sequence was developed that used the TINDERS system to record SWIR video of all 10 subjects in the second dataset with specific pose angles and facial expressions. The video was then post-processed for image extraction.

The data was collected in the outdoor parking structure of the WVHTCF facility. The TINDERS system was set up at one end of the structure. One seat was placed at a range of 70 meters from the system and another at a range of 160 meters at the other end of the structure. At both locations, back drops were placed behind the subjects to minimize any variability in background. The parking garage is dimly lit offering very little ambient lighting. Several chalk markings were strategically placed in front of both seats that forced the subjects to turn their heads approximately 20 degrees when looking at each one.

### 3.3 Image Acquisition

After the visible mug shot was acquired indoors, each subject was ushered to the 70 meter mark in the parking structure. The subject was instructed to stare with a neutral expression directly into the TINDERS system. At this point, the system was already running, had the illuminator on, had the imager zoom set to the minimum field of view (FOV), and was recording live video. The subject was asked to maintain this position and expression for 10 seconds, allowing for the recording of ~300 video frames at 30 frames per second. The subject was then instructed to look at the bottom left mark and say out loud, “One-one thousand, two-one thousand.” The subject was then instructed to look at the bottom center mark and repeat. In total there were nine marks that gave a complete angular view of the face. Having the subject speak out loud introduced variability into the facial expression recorded from frame to frame. Once this was completed, the subject was instructed to repeat this progression without speaking the words out loud, maintaining a constant, neutral expression. Example imagery from 70-m can be seen in Figures 4 and 5. The subject was then asked to proceed to the 160 meter mark.

At this point, the three exercises are repeated: ten second frontal neutral expression, talking angular progression, and neutral angular progression. This completes the acquisition process.

![Figure 4. High Correlation. SWIR Facial Imagery at 70-m range showing high correlation between images to be fused for facial recognition.](image)

![Figure 5. Low Correlation. SWIR Facial Imagery at 70-m range showing low correlation between images to be fused for facial recognition.](image)

Several aspects of the experimental conditions resulted in less than optimal imagery. First, some of the 70-m imagery and all of the 160-m imagery was inadvertently overexposed, with the effect most pronounced at 160-m. Figure 6 shows examples of overexposed imagery at 70-m and 160-m. In addition, high winds in the parking structure resulted in some image blurring due to camera motion, particularly at 160 m. Due to time constraints and harsh winter weather, the data collection was not repeated. While the overall face matching performance on this dataset is expected to be far lower
than typical TINDERS performance at these distances, the relative performance between high-correlation and low-
correlation fused imagery should still illustrate the effect of image correlation.

Figure 6. Image overexposure. Examples shown from distances of both 70 m (left) and 160 m (right).

Earlier internal WVHTCF experiments have indicated that matching performance of facial imagery at pose angles of 10-
degrees is comparable to that of frontal imagery; however, matching performance of facial imagery with 20-degree pose
angles is far lower. Thus, any positive effect observed on fused matching scores resulting from the fusion of 20-degree
facial imagery with frontal facial imagery may be significant.

3.4 Processing for Facial Recognition

Once the video was acquired, probe images were extracted for each subject corresponding to each of the different
imaging conditions. At both distances, a large number of highly-correlated frontal neutral-expression probe images were
extracted as well as each of the 9 angular poses with both neutral and talking expressions. Once the probe images were
extracted from the recorded video, they were grouped into several sets with high and low correlation. For each low-
correlation (high variability) set, there was a high-correlation set containing the same number of probe images. These
sets are listed in Table 1.

<table>
<thead>
<tr>
<th>70 Meters</th>
<th>160 Meters</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 frontal neutral</td>
<td>6 frontal neutral</td>
</tr>
<tr>
<td>15 frontal neutral</td>
<td>24 frontal neutral</td>
</tr>
<tr>
<td>24 frontal neutral</td>
<td>6 frontal neutral, 9 neutral angle progression, and 9 talking angle progression (24 total)</td>
</tr>
<tr>
<td>6 frontal neutral, 9 neutral angle progression (15 total)</td>
<td></td>
</tr>
<tr>
<td>6 frontal neutral, 9 neutral angle progression, and 9 talking angle progression (24 total)</td>
<td></td>
</tr>
</tbody>
</table>

Each set of probe images was then submitted to the TINDERS face recognition software for matching against the 114-
person visible-spectrum gallery, and the fused matching scores for the top 10 ranked candidates were displayed. For all
probe images, eye locations were marked manually. Because the matching scores depend upon the marked eye
locations, it was important to ensure that any probe image included in more than one set (e.g. the images in the 6 frontal
neutral set were included in all sets at the same distance) had the same marked eye positions in each set. For the 160-m
images, where severe overexposure led to low matching performance, the sensitivity of the results to the choice of eye
location required that the entire processing chain be performed 4 times, independently, with eye-locations independently
marked each time.

For each trial, the rank and fused score obtained by the subject of interest was recorded in a spreadsheet. This was done
for all ten test subjects and once for each of the 70-m probe sets and 4 times for each of the 160-m probe sets listed in the
Table above.
3.5 Cumulative Match Characteristics

The cumulative match characteristic (CMC) curve is a convenient tool to compare the face recognition accuracies resulting from the fusion of the different sets of probe images. Because the CMC curve directly shows how often a genuine match is highly ranked by the system, it is better suited to the current experiment than other types of metrics, such as the receiver operating characteristic. Overlaying the CMC curves for different sets of probe images will allow for easy visualization of the relative matching performance for the fusion of the different sets.

To calculate a CMC curve, the fraction of subjects correctly ranked better than or equal to each rank is computed. For instance the Rank 1 value of the CMC is just the fraction of subjects who were correctly identified at Rank 1. For Rank 2, the fraction of subjects correctly identified at Rank 2 or higher is computed, etc. Equation 1 provides the general formula for the CMC, where \( P(r) \) is the fraction of test subjects whose fused gallery image matching score had a rank of \( r \) out of all gallery images in the database, and \( k \) is the rank coordinate on the horizontal axis of the CMC curve.

\[
CMC(k) = \sum_{r=1}^{k} P(r); \quad k = 1, ..., m,
\]

3.6 Results for 70-m data

The CMC results for the high-correlation probe sets are shown in Figure 7. The left plot shows the baseline CMC for the fusion of a set of 6 frontal, neutral-expression facial images like those in Figure 4, while the right plot overlays the CMC curves for similar sets with 15 and 24 images. The Rank 1 performance for all three sets is 40%, while the Rank 10 performance of the 15 and 24-image set is 80% as compared to 70% for the 6-image set. This is consistent with a possible small benefit to fusing results for a larger number of very similar probe images, although this result is within the statistical error of the experiment. Note that with a database size of 114 faces, the purely random chance of guessing a Rank 1 match would be 0.9% and the chance of guessing Rank 10 or better would be ~ 9%.

Figure 8 compares the CMC curves for low-correlation probe sets to high-correlation probe sets acquired at 70-m range. The left plot shows the CMC for the 70-m set that includes 6 neutral-frontal images and 9 neutral angle progression images, while the right plot shows the set that also includes the 9 talking angle progression images. Both plots overlay CMC curves for high-correlation probe sets for comparison. In both cases, a clear performance improvement is seen when the fused dataset includes both neutral frontal and a progression of angled poses. In particular, both of the low-correlation image sets have 90% rank 4 performance and 100% rank 10 performance, while the high-correlation image sets have no better than 60% rank 4 performance and 80% rank 10 performance. This is despite the fact that the low-correlation SWIR probe image sets fuse the original 6 neutral-frontal images with additional SWIR images that have a 20º pose angle, which typically have low matching performance against frontal neutral visible gallery images.
Figure 8. Comparison of low-correlation to high-correlation CMC curves at 70-meters range. (left) CMC curve for set containing 6 neutral frontal and 9 neutral angle progression images overlaid with CMC curves for 6 and 15 neutral frontal images. (right) CMC curve for set containing 6 neutral frontal, 9 neutral angle progression, and 9 talking angle progression images overlaid with CMC curves for 6 and 24 neutral frontal images.

3.7 Results for 160-m data

As mentioned above, the quality of the 160-m images was severely degraded by overexposure and high winds, reducing the face matching performance well below typical TINDERS performance at this distance. When image quality is poor, face matching results are very sensitive to the marked eye locations, so 4 analysis groups of CMC curves were generated, each group having independent, but internally-consistent eye locations. Figure 9 shows the CMC curves for each of the four groups. Each plot includes a CMC for the 6 and 24 neutral-frontal probe image set as well as the 24-image set containing 6 neutral-frontal, 9 neutral angle progression, and 9 talking angle progression images.

Figure 9. Comparison of low-correlation (green) to high-correlation (blue and red) CMC curves at 160-meters range. The low-correlation probe set includes 6 neutral-frontal images, 9 neutral angle progression, and 9 talking angle progression images. High-correlation probe sets include 6 and 24 neutral-frontal images. Image quality was degraded by overexposure and motion blur due to high winds. Each plot uses the same images but with independent manual marking of eye locations. Note the large variability in the data between plots.
As can be seen from Figure 9, there was a large amount of statistical variability between the four analysis groups, and it was not possible to discern a statistically significant difference in matching performance between the high-correlation and low-correlation probe image sets at 160 m. While the degraded imagery led to overall low matching performance, achieving only 10% - 20% rank 1 success, and only 50% - 60% rank 10 success, the success rates were still far better than random chance, which would have predicted only 0.9% and 9% rank 1 and rank 10 success rates, respectively, indicating that the facial features still played an important role in determining the CMC curves. One possible explanation for the lack of observed benefit for fusing low-correlation images is the large 20º pose angle used in the angle progression images.

4. DISCUSSION

By fusing the face matching results from multiple probe images of the same person, it is possible to improve overall matching performance. It is logical to expect that probe images that closely resemble each other will produce similar matching scores against the gallery, and thus derive limited overall benefit from fusion. In contrast, it is logical to expect that probe images with varying pose angle and facial expression, within some useable range, will produce more variation in matching scores, thus deriving a larger benefit from fusion. It is also clear that probe images with very high pose angle or extreme expressions, where single-probe matching performance is extremely low, would not be expected to improve matching performance when fused with other probe images within the useable range. The 70-m results shown in Figure 8 clearly show that the fusion of 24 images with variable pose angles and expressions results in higher matching performance than the fusion of 24 images with very similar pose angle and expression. These results indicate that even images with pose angle of 20º can have a positive impact on matching performance, when the image quality is sufficiently high. In contrast, the 160-m results shown in Figure 9, which used low-quality, overexposed probe images, showed no statistically-significant improvement to matching performance when images with varying 20º poses were added to the fused probe image set. One possible explanation for this is that the range of useable pose angles decreases as image quality decreases. It is possible that a low-correlation probe image set of the same quality, using a smaller pose angle such as 10º, might have resulted in a larger improvement to the fused matching score at 160 m. Experiments that include a larger number of subjects, automatic eye location, and more low-to-intermediate pose angles will be needed to better understand and quantify the relationship between image correlation and fused face matching performance.

5. ACKNOWLEDGMENTS

This research was performed under contract N00014-09-C-0064 from the Office of Naval Research, with funding and oversight from the Deployable Force Protection Science and Technology Program. The authors would like to acknowledge important technical contributions from Jason Stanley, Kenneth Witt, William McCormick, and the MorphoTrust USA team, as well as the cooperation of the WVU Center for Identification Technology Research.

REFERENCES