Multi-kernel aggregation of local and global features in long-wave infrared for detection of SWAT teams in challenging environments

Ankit S. Arya\textsuperscript{a}, Derek T. Anderson\textsuperscript{b}, Cindy L. Bethel\textsuperscript{a,c}, Daniel Carruth\textsuperscript{c}

\textsuperscript{a}Computer Science and Engineering, \textsuperscript{b}Electrical and Computer Engineering, \textsuperscript{c}Center for Advanced Vehicular Systems, Mississippi State University, Mississippi State, MS, 39759.

ABSTRACT

A vision system was designed for people detection to provide support to SWAT team members operating in challenging environments such as low-to-no light, smoke, etc. When the vision system is mounted on a mobile robot platform: it will enable the robot to function as an effective member of the SWAT team; to provide surveillance information; to make first contact with suspects; and provide safe entry for team members. The vision task is challenging because SWAT team members are typically concealed, carry various equipment such as shields, and perform tactical and stealthy maneuvers. Occlusion is a particular challenge because team members operate in close proximity to one another. An uncooled electro-optical/long wave infrared (EO/LWIR) camera, 7.5 to 13.5 m, was used. A unique thermal dataset was collected of SWAT team members from multiple teams performing tactical maneuvers during monthly training exercises. Our approach consisted of two stages: an object detector trained on people to find candidate windows, and a secondary feature extraction, multi-kernel (MK) aggregation and classification step to distinguish between SWAT team members and civilians. Two types of thermal features, local and global, are presented based on maximally stable extremal region (MSER) blob detection. Support vector machine (SVM) classification results of approximately [70, 93]\% for SWAT team member detection are reported based on the exploration of different combinations of visual information in terms of training data.

Keywords: electro-optical, long wave infrared, EO/LWIR, SWAT, people detection, human robot interaction

1. INTRODUCTION

Law enforcement involves inherently dangerous situations while operating under visually challenging conditions such as low-to-no light and smoke or gas-filled environments. Law enforcement officers put themselves in harm’s way on a consistent basis to keep others safe. The use of robotic systems as teammates can potentially reduce the risks to officers, civilians, and even offenders. Law enforcement officers have expressed a desire to use robotic systems as a force multiplier by providing them with new capabilities through the use of a robotic teammate. A current issue with the use of robots in law enforcement applications is that an officer must be taken out of the fight to teleoperate the robot. This reduces the overall effectiveness of the team and can be more of a distraction than a help. For a robot to be effective in the role of a teammate, it must interact with the team through the use of limited supervisory control and utilize detection methods necessary to interpret commonly used law enforcement and tactical team commands to be considered a more independent team member.

When a law enforcement officer is called to investigate an incident, they must quickly assess the situation and environment. Law enforcement officers may be called in to respond to late night alarms operating under low-to-no light conditions (e.g., banks, warehouses, gun shops, etc.). This type of investigation requires the officers to systematically and exhaustively explore the location in visually challenging environments. This slow and methodical search can be quite challenging, and it would be preferred to have the assistance of a robot that could explore the areas in advance of a law enforcement officer entering an environment. A robot equipped with thermal and night vision capabilities would be able to detect if there is a human present that may have activated the alarm system. The vision system would be capable of sending this information back to the law enforcement officer that may be located outside of the building. This will provide invaluable information to
the officer regarding any obstacles present in the environment, humans present, and other relevant information that may not be easily detected with visual spectrum cameras or even with human vision systems. This type of imaging feedback will make entry safer for law enforcement officers.

Tactical teams face even greater dangers in their responses. They typically operate in a high stress environment, often involving drugs, alcohol, and/or hostages. It is often the case that special weapons and tactics (SWAT) team members must enter into dangerous and unknown environments. They do not know what awaits them when they enter into a dwelling or building. At times they will use smoke to attempt to get suspects to surrender and exit these dwellings; however if they do not exit, then the team members must enter not only a dark environment but also a smoke-filled environment which presents many visual challenges. A robot can be sent into this type of environment as a distraction, but also as a means of collecting visual information through the use of thermal imaging to aid SWAT team members with identifying the location of humans in the dwelling as well as animals. It may even be possible through thermal imaging to identify weapons that may impact the response. Using a robot to make first contact and also to gather information in visually challenging environments can save officer lives as well as the lives of others that may be present in the environment.

Fig. 1(a) is the robot platform used in this article (Clearpath Husky A200). The robot platform was equipped with multiple cameras, such as the Microsoft Kinect and a long wave infrared (LWIR) camera. Fig. 1(b) is an example of the robot in training exercise with the SWAT team for a slow and methodical clearance task.

2. BACKGROUND

Numerous advancements have been made in the area of computer vision in the last two decades. In part, the fields of pattern recognition and machine learning have played a significant role in many of these breakthroughs. Another important thrust is advancements in technology (e.g., cameras, smart phones, social media, etc.). While computers are still not able to answer many deep image understanding questions, in comparison to a human, some sub-topics have seen great advances. For example, the topic of object detection and classification has made significant progress. While some designs vary specific detectors for humans, others have cast the topic of human detection as a sub-object detection problem. That is, the design of generic object detection algorithms combined with a collection of human training data. The PASCAL VOC challenge\(^1\) is a sort of “standard” in the vision community. It provides a collection of vision techniques and their performance on different object categories. Noteworthy vision techniques include; poslets,\(^2\) deformable part models,\(^3\) object detector with boosting\(^4\) and the “fastest detector in the west”.\(^5\)

Herein, we used the discriminatively trained deformable part models by Pedro F. Felzenszwalb.\(^3\) This detector can be trained on people, it is a top performer in terms of both run-time and people detection, and the authors have also made source code available for academic use. Note, computer vision for SWAT detection on a robot is
very much real-time constrained. Specifically, one desires a solution that has minimal computational complexity for an accuracy specified by a given domain/task. In this work, the object detector is used simply as a pre-screener. That is, it is trained to recognize all types of people, e.g., civilian and SWAT. It is the role of subsequent processing to identify SWAT (friend) from civilian (e.g., foe). Felzenszwalb’s part-based method uses a histogram of gradients (HOG) to construct strong low-level features. An object is represented as a collection of parts in deformable configuration (which is modeled separately). Also, discriminative learning with latent variables is performed using latent SVM.

Another noteworthy computer vision approach is the so-called topic of deep learning. Solutions can be categorized into single layer, multi-layer perceptron or classical bag of words (BOWs). A deep architecture is capable of learning high-level features using multiple layers that build on-top of each other. The complaint of deep learning researchers is that “standard CV” pipelines are “hand crafted”, i.e., built on top of SIFT, HOG, GLOH, textons, vector quantization, spatial pooling, SMVs, etc. In addition to these two different approaches, there is another hybrid approach that utilizes a combination of both methods. Yang’s method replaced vector quantization with sparse coding for codebook learning. Also, they used dense SIFT features instead of raw image patches for sparsity coding. A linear classifier was used thereafter to perform classification. However, the application-based drawback is that it is sensor dependent. Therefore, one must learn and extract sensor specific features from the robot.

3. METHODS

In this section, a new computer vision (CV) algorithm is put forth for SWAT team member detection. Figure 2 is an illustration of the proposed framework and Algorithm 1 is its corresponding high-level algorithmic description. The remaining sub-sections detail the different computational steps outlined in Algorithm 1.

As Figure 2 indicates, the robot first collected imagery using an uncooled EO/LWIR camera. The next step was people detection. A goal of this project was to identify both SWAT team members and civilians, which is
Algorithm 1 Proposed SWAT team member detection algorithm in EO/LWIR.

1: Run deformable part models algorithm \(^3\) on image \(I_t\) (1 ≤ \(t\) ≤ \(T\)) to generate a set of training candidate windows, \(I_n(x, y)\) (1 ≤ \(n\) ≤ \(N\)).
2: Perform median filter (3x3 window) on \(I_n\) to reduce noise while preserving edges.
3: Extract (light-on-dark) MSERs, \(r_n = MSER(I_n)\), where \(r_{n,k}\) is the set of pixel coordinates in blob \(k\).
4: Build MSER aggregate image using Algorithm 4
5: Extract local features from keypoints (MSERs) using Algorithm 2
6: Training: use labeled data and run sparse coding \(^7\) on local features to produce codebook \(C\) of size 128.
7: Build BOW descriptor using codebook \(C\) and MSER local features (see Algorithm 3).
8: Build global LWIR feature using Algorithm 5.
9: Perform multi-kernel aggregation and classification.

needed to ultimately determine friend versus foe. However, the task of determining friend versus foe is beyond the scope of this work as it likely involves a deeper understanding of the scene and it will also likely require activity/behavior analyses. Herein, a *general purpose* person detector was trained and used as a pre-screener to find all evidence of people. The goal of a pre-screener is to achieve the highest possible positive detection rate (PDR) (e.g., 100%) and to minimize the system’s false alarm rate (FAR). However, this approach typically results in an unacceptable FAR, but one that is still better than a brute force search. The result of the pre-screener was a set of candidate windows. A candidate window is a rectangular sub-image region, typically an *axis aligned bounding box* (AABB), that contains as much of the object (e.g., human) and as little of the surrounding background as possible. Note, the CV system put forth here was not restricted to a specific type of person detector. Any algorithm that produces (or can be made to produce) AABBs can be utilized. This is important as people detection is a fast moving field and algorithmic approaches vary greatly. Herein, the deformable parts model algorithm \(^3\) was used as its reported performance (accuracy) is relatively high, its computational complexity is relatively low, and the source code is freely available. It was the objective of the later steps in our CV pipeline (Figure 2) to refine the initial people detection results and identify a particular class of people (e.g., SWAT).

After candidate windows were identified, keypoint detection was performed using the MSER algorithm. Local features were extracted from keypoints and aggregated into a *bag of (visual) words* (BOW) descriptor. This descriptor characterizes a candidate window according to its local texture and shape information. For example, local features specific to specialized equipment and weapons can be used to help identify a SWAT team member. In addition, light-on-dark “blobs” that may represent regions that are hotter than their surroundings in LWIR, are found using the MSER algorithm. A new global feature was put forth that characterized the relative distribution of “heat” within a candidate window. This feature helped to refine the candidate search space and provided better discrimination between SWAT team members and civilians. In the final step, this local and global information was combined (per-candidate window) and it was subjected to multiple kernel (MK) *support vector machine*- (SVM) based classification. The remainder of this section describes these different CV stages.

3.1 Long wave infrared camera

The use of low lux and thermal imaging technologies is not new to CV, they are used for tasks such as surveillance, \(^8\) *automatic target recognition* (ATR), \(^9\) and *explosive hazard detection* (EHD). \(^10\) For example, most modern *unmanned aerial vehicles* (UAVs) are equipped with an EO/IR gimbal camera system for *intelligence, surveillance, and reconnaissance* (ISR). Of particular interest to this project is the so-called *long wave infrared* (LWIR) spectrum. LWIR can generate quality imagery in challenging environments such as low-to-no light, and smoke- or gas-filled environments in which SWAT teams typically operate. LWIR is significant because it provides enhanced capabilities, compared to a human. However, LWIR is not without limitations and challenges. LWIR is impacted by a restricted range in environmental conditions such as moderate-to-high moisture (e.g., rain and fog). The exact range depends on a number of factors, such as atmospheric conditions, type of fog, type of IR camera used, and the properties of the targets (e.g., size, temperature difference of target to its background, etc.). In particular, passive LWIR sensors have an additional benefit that they do not require an additional illumination source that could draw unwanted attention in a SWAT setting. Also, LWIR sensors generally help
simplify the human detection process. Unlike visual spectrum imagery (i.e., standard RGB CCD cameras), challenging texture, shape, and other non-human image information is generally not as abundant (example in Figure 3).

Two main options exist for LWIR: cooled and uncooled. Uncooled thermal imagers have the advantage of lower cost and mobility, compared to a cryogenic-cooled core, making it a viable option for SWAT applications. A slight drawback of uncooled cores is that they measure relative thermal contrast in a scene and they tend to have a limited (e.g., 8-bit) range (pixel depth). However, for our application the pros outweighed the cons. In this article, an uncooled LWIR FLIR tau 640 camera with a 9mm lens was used. Figure 3 (left) is an example LWIR image from the camera.

3.2 Maximally stable extremal regions

As mentioned, the result of our pre-screener (object detector trained on people) was a set of candidate windows for each image. The goal of the latter processing stages, was to extract features that helped distinguish SWAT from non-SWAT. The first step was keypoint detection, also referred to as a saliency detector, interesting point detector, etc. A large number of algorithms exist for this task, such as David Lowe’s difference of gaussians (DoGs), the Harris detector, the dual tree complex Wavelet transform, etc. Most keypoint detectors, e.g., DoG, are good at detecting features such as corners. However, the goal of this work was to segment different regions of the body (or “blobs”) and to measure local features and a per-candidate window global heat feature. The MSER algorithm was selected as it has been successfully used in recognition, matching, and tracking applications.

Specifically, the MSER was used herein to extract “bright-on-dark” blobs from a candidate window, $I_n$. The freely available VLFEAT MSER implementation was used. The result was the extraction of multiple blobs at different translations, rotations, and scales. However, the MSER algorithm is not guaranteed to only return human segments (i.e., arms, legs, faces, etc). The MSER is a relatively low-level image processing operator that finds all blobs that pass a user specified criteria (set of conditions for being a MSER). It was the job of the local features, global features, and pattern analysis (classifier) to distinguish between SWAT and non-SWAT.

3.3 Local features

It is asserted here that two major types of evidence for SWAT team member classification exists in LWIR and should be exploited. In this sub-section the first source of evidence, local features, is discussed. Local features are critical for detecting specific visual cues present in the thermal imagery. SWAT examples include helmets, shields, protective vests, and weapons. Algorithm 2 describes the steps involved in the calculation of local features.
Algorithm 2 Extraction of local features in a candidate window. See Section 3.3 for algorithmic implementation details, e.g., parameters.

1: Initialize the matrix $L_{I_n}$ to 0  \hspace{1em} \triangleright \text{Matrix of all 0's of size } |r_n| \times 128
2: for $k = 1$ to $|r_n|$ do  \hspace{1em} \triangleright \text{Process all MSERs}
3: Find minimum area rectangle that encapsulates $r_{n,k}$
4: Resample bounded region to produce $b_k$ of size 25x25  \hspace{1em} \triangleright \text{Use bi-linear interpolation}
5: $l_k = \text{HOG}(b_k)$  \hspace{1em} \triangleright \text{Extract cell-structured HOGs}
6: $L_{I_n}(k, 1 : 128) = l_k$  \hspace{1em} \triangleright \text{Add feature to matrix}
7: end for

The goal of Algorithm 2 is to extract the histogram of gradients (HOGs) per MSER blob in a candidate window. As the name implies, the HOG descriptor is the aggregation of local gradient information (i.e., texture) in a window into a histogram. Gradient orientation information was used to index the histogram and gradient magnitudes were the “amount of contribution” added to the bins. Specifically, an alpha interpolation in the closest and next closest bins were used per pixel. First, an AABB for each MSER blob, $r_{n,k}$, was calculated. Next, that AABB was re-sampled, using bi-linear interpolation, to an image chip, $b_k$, of size 25x25. It was important that a consistent procedure was used to calculate the gradients (i.e., they were at a similar scale) and populated the HOG (i.e., one histogram was not populated using a drastically different sampling amount). The procedure described by David Lowe\cite{11} was used for the HOG calculation herein. Namely, a 4x4 cell-structured descriptor was produced. Cell-structured descriptors preserved some of the spatial context of texture in a local region. Each cell had a histogram of 8 bins. Therefore, each bin spanned 45 degrees. A 4x4 descriptor has 16 bins and their concatenation is of size $16 \times 8 = 128$. Each MSER resulted in a single cell-structured HOG.

A common CV approach is to combine a set of local features into a higher-level feature for region/scene-level characterization. The BOW approach was selected to combine all MSER-based cell-structured HOGs for a candidate window. Each “word” in a BOW is typically a local visual descriptor (e.g., HOG, LBP, etc.). For example, in the recognition of human faces different visual descriptors (words) may exist for landmarks such as eyes, nose, mouth, etc. The visual dictionary is typically created (learned) using a training dataset. A quantization or clustering algorithm is used to acquire a relatively small dictionary of visual words. Herein, the sparse coding algorithm by Honglak Lee\cite{7} was used. After the dictionary was learned, Algorithm 3 was used to build the BoWs to train a classifier.

Algorithm 3 Construction of BOWs.

1: Set $B = \phi$  \hspace{1em} \triangleright \text{initialize the BOW}
2: for $n = 1$ to $N$ do  \hspace{1em} \triangleright \text{for each candidate window}
3: $Y = L_{I_n} \ast C^T$  \hspace{1em} \triangleright C is codebook (size $P \times 128$)
4: $B_{I_n} = \max(Y)$  \hspace{1em} \triangleright \text{row-wise max (size } 1 \times P)
5: end for

In Algorithm 3, each local feature in a candidate window is compared to the dictionary and the closest word is selected. Note, this task can be (and was herein) implemented in a “soft” and efficient fashion using linear algebra. As Algorithm 2 shows, the first step is to create a matrix of $|r_n| \times 128$ sized HOG descriptors. This matrix was then multiplied by the codebook matrix, which was of size $1 \times P$ (where $P$ was the number of visual words). The result was a matrix of the dimensionality number of local features by $P$. A number of methods exist to reduce this matrix and produce a BOW (e.g., maximum, summation, etc.). Herein, the maximum was selected based on its performance reported in Yang’s paper on Linear spatial pyramid matching\cite{6}.

3.4 Global heat signature feature

In the last sub-section, the process of local feature extraction and BOW construction was outlined. The second source of information exploited herein is a (relative) global heat signature. It was observed that SWAT team members exhibited a different “heat signature” than civilians. Specifically, SWAT team members wear protective armor that covers significant portions of their body. Regardless of the specific type of armor worn, the heat
signature of a SWAT member appeared to vary between the armor covered and the exposed body parts, such as face, neck, and underarms. The armor covered region was not expressed as significantly in LWIR as other exposed areas with less gear. This resulted in more variance in the heat signature of a SWAT team member compared to a civilian. The information regarding the heat signature difference was extracted in the form of a global feature (global to each candidate window) and it was used alongside the local features to help improve target recognition. Note, this article assumes that SWAT team members primarily operate in a standing or squatting mobile pose during tactical maneuvers. Algorithm’s 4 and 5 are formal descriptions of the global heat feature extraction from a candidate window.

Algorithm 4 Extraction of MSER aggregate image for candidate window. See Section 3.2 for details.

1: Initialize matrix \( M_n \) to all 0's
2: for \( k = 1 \) to \( |r_n| \) do
   3: Calculate the median value, \( d_k \), of blob \( k \)
   4: for all \((i,j)\) in \( r_n,k \) do
      5: \( M_n(i,j) = \max(M_n(i,j),d_k) \)
   6: end for
7: end for

Algorithm 5 Extraction of global LWIR feature.

1: Set \( G_{I_n} \) to all 0's
2: \( C_n = \text{cov}(M_n) \)
3: Calculate the eigenvectors and eigenvalues of \( C_n \)
4: \( M_n = \text{rotate}(M_n,\theta) \)
5: \( G_{I_n}(k) = \text{mean}(M_n(k,\hat{U})) \)
6: end for

First, in Algorithm 4 all MSERs are calculated for a candidate window. Next, a new image chip is created, which is the same size as the candidate window created. For each MSER, the average LWIR value is recorded and it is used to update the aggregate image chip. Specifically, each pixel is assigned the maximum average MSER LWIR value. The resultant image is a soft segmentation of all light-on-dark blobs.

Second, in Algorithm 5, a covariance matrix is calculated on the MSER aggregate image \( M_n \) and the major axis is identified (i.e., the eigenvector with the largest corresponding eigenvalue). Note, the covariance matrix is a weighted (based on the MSER aggregate image values) pixel coordinate-wise calculation. The image chips are then rotated so that the direction of maximal spread is upwards (the image height direction). After rotation, the mean LWIR value for all non-zero entities along the horizontal axis of the MSER aggregate image is calculated. The set of mean values is the global (to a candidate window) LWIR feature vector. As each candidate window is of size 64x128, the resulting global feature vector is of length 1x128.

3.5 Classification

After local and global features were extracted, a decision must be made regarding if a candidate window is a member of the class SWAT. Herein, SVM-based classification was used. Specifically, we used the popular and efficient open-source LIBSVM implementation. We explored the utilization of these features in a number of regards. First, we investigated the benefit of using just one feature (i.e., local or global, but not both). Second, we investigated if there was any benefit in directly combining these features (i.e., feature vector concatenation). This was by far the simplest and most common approach. Third, we investigated if there was any benefit in aggregating the output of multiple classifiers (SVMs). Specifically, the local and global features were used individually to train separate SVMs based on different kernels (e.g., linear and radial basis function (RBF) kernels). The direct
approach can be considered a crude form of feature-level fusion, while the post-SVM approach is a multi-classifier aggregation.

In the case of multi-SVM output aggregation (the third option described above), two models were trained. Let $\Phi_1$ be the trained SVM using a linear kernel and local BOW features and let $\Phi_2$ be the trained model using an RBF kernel and global features. Note, the linear and RBF selections were determined experimentally herein. During testing, $\Phi_1$ and $\Phi_2$ were used to calculate the probability that a candidate window belongs to a particular class. Specifically, one acquires a probability per class (SWAT and non-SWAT) per-model. These probabilities are aggregated using

$$p_1 = \text{linear kernel}(\Phi_1)$$
$$p_2 = \text{rbf kernel}(\Phi_2)$$
$$CM = \max(|p_1(2) - p_1(1)|, |p_2(2) - p_2(1)|).$$

This post-classifier approach used the maximum operator over $p_1$ and $p_2$ (where $p_i(1) \in [0, 1]$ was the probability of class 1, SWAT, and $p_i(2) \in [0, 1]$ was the probability of class 2, non-SWAT) to calculate the final system confidence, $CM \in [0, 1]$.

4. DATASET

A unique thermal dataset of SWAT team members performing tactical maneuvers was collected. To the best of our knowledge, there is no publicly available equivalent to this dataset. The mere collection of this data represents a significant effort (cooperation) between local law enforcement and academia.

![Figure 4. Example imagery from our different datasets. D1(Upper Left), D2(Upper Right), D3(Lower Left), and D4(Lower Right).](image)

The general interest of our group is to research and develop assistive robotic and vision-based technologies for domain experts (law enforcement) so they may effectively and safely perform their duties. This dataset consists of four sub-datasets captured on different days and locations. Activities performed, camera viewpoint, time of day, location of recording, and other factors differ across the four datasets. The variability across the datasets captures the reality that a CV system will encounter when deployed in the field. These different (sub-)datasets are significant herein because they include variation in terms of visual content, backgrounds, participants, viewpoints,
and degrees of occlusion for the purpose of training and testing the generalizability of the proposed algorithm. Example imagery is shown in Figure 4.

Datasets 1 and 2 (D_1 and D_2) were recorded in one building in two different hallways. Dataset 3 (D_3) was collected in a different building than D_1 and D_2. Specifically, D_3 is unique because it has glass walls (which can be a challenge for LWIR). Dataset 4 (D_4) is significantly different from the other datasets as it was collected in an open area on a sunny day without SWAT. In addition, a limited amount of data was collected for the civilian class inside of a building while SWAT was performing their operations. As a result, our datasets include a large collection of SWAT imagery for training classifiers, however contain insufficient civilian imagery to train a civilian classifier. In general, civilian data is easier to obtain, but not in the context of SWAT operations. We are addressing that in future work. As a result, datasets D_1-D_3 have less in-class variation, as class SWAT is represented by 8 different officers and only 3 different civilians. This was the motivating factor for collecting D_4. It was decided to capture D_4 in an entirely different environment to further investigate the system’s generalizability. For example, most CV techniques suffer from “memorization” of a dataset (i.e., poor generalizability to new unknown scenes). In part, this can be attributed to feature extraction. That is, classifiers identify patterns based in part on features and most features extracted from AABBs capture both foreground (i.e., human) as well as background. Our approach is an attempt, in part, to circumvent some of these shortcomings. By extracting MSERs (blobs) in LWIR, our goal was to sense more human-centric information and place system focus (i.e., our local and global features) on sub-AABB image content that was most likely human, not background.

5. EXPERIMENTS, RESULTS AND DISCUSSION

In this section, the training and testing methodology used herein is described. First, candidate windows were generated for the entire dataset by the people detector. These images were then manually (by a human) labeled as SWAT or non-SWAT. Next, different combinations of the datasets, i.e., D_1-D_4, were used for training and testing. Specifically, each testing dataset was trained with the combination of other three datasets. For example, if the testing data was D_1 then the combination of training data was (D_2,D_3), (D_2,D_3,D_4), and (D_1,D_2,D_3). Since, each dataset was captured at a different time, contained different numbers of SWAT members and civilians, and was performed at different locations from different viewpoints. This testing approach was far less biased than a random sampling of the entire unioned dataset. That is, these datasets provided a better evaluation of how well the classifiers transferred from one environment to another.

At this point in time, parameters were empirically determined. In future work we will address the parameter learning for the entire system. With respect to the MSER algorithm, VLFEAT’s MSER Matlab implementation was used. The specific set of parameters include: ‘MaxArea’ 0.25, ‘MinArea’ 0.0005, ‘MinDiversity’ 0.1, ‘MaxVariation’ 0.7, ‘BrightOnDark’ 1, ‘DarkOnBright’ 0, and ‘Delta’ 5. In other words, a blob can be no larger than 25% and no smaller than 0.0005% of the candidate window size. The next two parameters were thresholds for determining the minimum inner blob diversity and variation. The next two parameters indicated that we only extract ‘bright-on-dark’ blobs, not ‘dark-on-bright’ MSERs (i.e., areas where the blob was hotter than its surroundings). Also, the delta value was set relatively low. The higher the delta value the stricter the blob extraction criteria (meaning higher delta values required greater contrast between a blob and its surroundings). We used a low delta value to identify subtle contrast changes inside the candidate window. With respect to codebook learning, we used Lee’s implementation of the sparsity coding algorithm with a codebook size of 128. Finally, the LIBSVM library was used for SVM classification.

Table 1 describes performance of the system on different training/testing data. Note, for the case of SVM probability aggregation, the optimal configuration discovered was the linear kernel for local features and the RBF kernel for the global features (with a weighting combination of 0.25 and 0.75 respectively). Overall, Table 1 tells the following story. The best, i.e., most consistent and highest accuracy, performer was the proposed post-SVM probability aggregation of kernels trained on separate features. That is, the best scenario involved the aggregation of multiple classifier outputs, each of which utilized different kernels. The system achieved a quality average accuracy rate between [70.01, 93.33] for the different challenging scenes put forth. An advantage of the post-SVM aggregation was that our algorithm was not dependent on just one feature and kernel per say.
by aggregating its result with the other kernel. In addition, it was observed that the RBF kernel trained on global features outperformed, on average, the linear kernel trained on local BOW features. It is interesting to note that the RBF kernel, which is usually less generalizable, was in our case, the best performer when trained using global features even on significantly different testing data. However, as we showed, the combination of the two leads to a more robust and higher accuracy solution.

### 6. CONCLUSION AND FUTURE WORK

In summary, we put forth a new approach to SWAT team member detection from LWIR for human-robot interaction. Namely, we proposed a new global LWIR feature, local BOW feature, and demonstrated that a post-SVM multiple kernel approach led to a more robust and higher accuracy CV solution. In future work, we will collect more civilian thermal data in conjunction with SWAT. We will also explore low-level multi-kernel fusion (at the kernel-level) for more intelligent feature space fusion. The system parameters were empirically determined (e.g., MSER parameters, post-SVM aggregation weights, etc). Future work includes learning these parameters.

### REFERENCES


