Proceedings of the 2007 IEEE/RSJ International
Conference on Intelligent Robots and Systems
San Diego, CA, USA, Oct 29 - Nov 2, 2007

Genetic MRF model optimization for real-time victim detection in Search and Rescue

Alexander Kleiner and Rainer Kümmerle
Department of Computer Science
University of Freiburg
D-79110 Freiburg, Germany
{kleiner,kuemmerl}@informatik.uni-freiburg.de

***Abstract*— One primary goal in rescue robotics is to deploy a team of robots for coordinated victim search after a disaster. This requires robots to perform subtasks, such as victim detection, in real-time. Human detection by computationally cheap techniques, such as color thresholding, turn out to produce a large number of *false-positives*. Markov Random Fields (MRFs) can be utilized to combine the local evidence of multiple weak classifiers in order to improve the detection rate. However, inference in MRFs is computational expensive.**

In this paper we present a novel approach for the genetic optimizing of the building process of MRF models. The genetic algorithm determines offline relevant neighborhood relations with respect to the data, which are then utilized for generating efficient MRF models from video streams during runtime.

Experimental results clearly show that compared to a Support Vector Machine (SVM) based classifier, the optimized MRF models significantly reduce the false-positive rate. Furthermore, the optimized models turned out to be up to five times faster then the non-optimized ones at nearly the same detection rate.

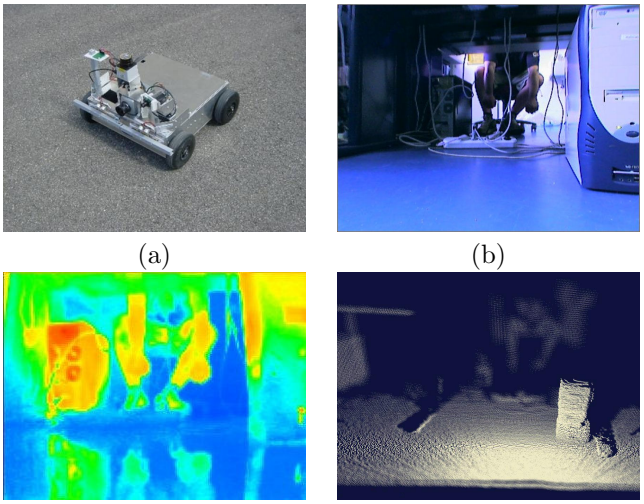


Fig. 1. The autonomous rescue robot *Zerg* (a) and vision data from the same scene (b)-(d): A color image taken by the CCD camera (b), a thermo image taken by the IR camera (c), and a 3D point set taken by the 3D scanner (d).

I. INTRODUCTION

The increasing extent of natural disasters, particularly earth quakes, hurricanes, and tsunamis, motivates research in the field of rescue robotics. One primary goal is to deploy, under the surveillance of a human operator, a team of robots for coordinated victim search after a disaster. This requires robots to perform subtasks, such as victim detection, partially or even fully autonomous.

The National Institute of Standards and Technology (NIST) develops test arenas for the simulation of situations after a disaster [10]. In this real-time scenario, robots have to explore an unknown area autonomously within 20 minutes, and to detect victims therein. There might be "faked" victim evidence, such as printed images of human faces, non-human motion, and heat sources that do not correspond to victims. The PC depicted in Figure 1 (b), for example, would be wrongly reported as a victim by most heat-seeking robots. Note that this example is particularly difficult due to the large size of the thermo signature (see Figure 1 (c)), as well as the nearby evidences caused by the presence of a true victim.

Due to the real-time constraint in rescue-like applications, only fast computable techniques are admissible. We successfully applied *color thresholding*, *motion detection*, and *shape detection* on images taken by an infrared and color camera[1]. However, the detection rate of these classifiers turns out to be moderate, since they are typically tuned for specific objects found in the environment. Hence, in environments containing many diverse objects, they tend to produce a large number of evidence frames, from which in the worst case, most are *false-positives*, i.e., objects that are wrongly recognized as victims.

One solution to this problem is to combine local evidences, i.e., evidences that are close to each other in the real world, and to reason on their true class label with respect to their neighborhood relations. Markov Random Fields (MRFs) provides a probabilistic framework for representing such local dependencies. However, inference

[1]The underlying vision system was part of the *Rescue Robots Freiburg* team, which won the 1st price of the autonomy rescue competition during RoboCup '05 and RoboCup '06.

1-4244-0912-8/07/$25.00 ©2007 IEEE.

in MRFs is computational expensive, and hence not generally applicable in real-time.

In this paper, we present a novel approach for the genetic optimization of the building process of MRF models. The genetic algorithm determines offline relevant neighborhood relations, for example the relevance of the relation between evidence types heat and motion, with respect to the data. These pre-selected types are then utilized for generating MRF models during runtime. First, the vertices of the MRF graph are constructed from the output of the weak classifiers. Second, edges between these nodes are added if the specific type of nodes can be connected by an edge type that has been selected during the optimization procedure.

Experiments carried out on test data generated in environments of the NIST benchmark clearly show that compared to a Support Vector Machine (SVM) based classifier, the optimized MRF models significantly reduce the false-positive rate. Furthermore, the optimized models turned out to be up to five times faster than the non-optimized ones at nearly the same detection rate.

Human body detection and tracking from color images has been already successfully applied based on background subtraction [18], [9], [3], and based on color thresholding [7]. SVMs have been utilized to detect human motion [6], [13], and MRFs have been applied for pedestrian tracking [19] and face detection [8]. In the context of rescue robotics, Bahadori et al. studied various techniques from computer vision and their applicability to the rescue context [2]. Nourbakhsh et al. utilized a sensor fusion approach for incorporating the measurements from a microphone, IR camera, and conventional CCD camera [12]. They assigned to each sensor a confidence value indicating the certainty of measurements from this sensor and calculated the probability of human presence by summing over all single sensor observation probabilities, weighted by their confidence value.

The remainder of this paper is structured as follows. In Section II we introduce the underlying vision system, Section III explains the MRF model, and in Section IV we introduce the genetic model selection approach. Finally, we provide results from experiments in Section V and conclude in Section VI.

II. VISION DATA PROCESSING

The utilized vision system is part of the rescue robot *Zerg*, shown in Figure 1 (a), which is equipped with a *Hokuyo* URG-X004 laser range finder (LRF), a *Thermal-Eye* infrared (IR) camera, and a *Sony DFW-V500* color camera. The LRF is capable of measuring distances up to 4000 mm within a field of view (FOV) of 240°, whereas the FOV of the IR and color camera are 50° and 70°, respectively. In order to combine evidence from thermo and color images, we firstly project their pixels onto the 3D range scan, and secondly determine pixel-to-pixel correspondence by interpolating from best matching yaw and pitch angles found in both projections.

Beforehand, camera images are linearized with respect to the intrinsic parameters of the camera system. On color cameras, these parameters are usually calibrated from pixel-to-real-world correspondences generated by a test pattern, such as the printout of a chess board [4]. In case of IR camera calibration, it is necessary to generate a test pattern that also appears on thermo images. This has been achieved by taking images from a heat reflecting metal plate covered with quadratic isolation patches in a chess board-like manner.

From both images three different evidence types are generated, which are *color*, *motion*, and *shape*, respectively. Each evidence type is represented by a rectangular region described by the position and size (u, v, w, h) on the image, number of pixels included, and the real world position (x, y, z) of the center.

Color pixels are segmented by fast thresholding [5] in the YUV color space. In case of the IR camera, only the luminance (brightness) channel is used since thermo images are represented by single values proportional to the detected temperature. Pixels within the same color class are merged into *blobs* by run length encoding, and represented by rectangular regions.

Motion is detected by background subtraction of subsequent images. Let I_t be an image at time t from a sequence of images $\mathbf{I}$ with $I_0 = Background$. Then, the difference between an image and the background can be calculated by $AVG_t = (1-\beta)\, AVG_{t-1} + \beta I_t$, $DIFF_t = AVG_t - I_t$, where AVG is the running average over all images and β a factor controlling the trade-off between latency and robustness. Finally, extracted foreground pixels are also merged into groupings by run length encoding and are represented by a set of rectangular regions.

Shape detection is currently limited to the detection of human faces. We use the *openCV* [4] implementation of the method from Viola et al. [17], which has been further improved by Lienhart [11]. The method utilizes a cascade of *haar*-like features that are trained and boosted from hundreds of sample images scaled to the same size.

III. Markov Random Fields

The series of images in Figure 4(a) clearly shows that single evidences are not sufficient to uniquely identify victims. Therefore, it is necessary to consider neighborhood relations in order to reduce false-positive detections. Markov Random Fields (MRFs) provides a probabilistic framework for representing local dependencies. A MRF is defined by an undirected graph $\mathcal{G} = (\mathrm{Y}, \mathcal{E})$, where Y is a set of discrete variables $\mathrm{Y} = \{Y_1, \ldots, Y_N\}$, and $\mathcal{E}$ is a set of edges between them. Each variable $Y_i \in \{1, \ldots, K\}$ can take on one of K possible states. Hence, $\mathcal{G}$ describes a joint distribution over $\{1, \ldots, K\}^N$.

According to the approach of Anguelov et al. [1], we utilize *pairwise* Markov networks, where a potential $\phi(y_i)$ is associated to each node and a potential $\phi(y_i, y_j)$, to each undirected edge $\mathcal{E} = \{(ij)\}\,(i < j)$ between two

nodes. Consequently, the pairwise MRF model represents the joint distribution by:

$$P_\phi(\mathrm{y}) = \frac{1}{Z}\prod_{i=1}^{N}\phi_i(y_i)\prod_{(ij)\in\mathcal{E}}\phi_{ij}(y_i, y_j), \qquad (1)$$

where Z denotes a normalization constant, given by $Z = \sum_{\mathrm{y}'}\prod_{i=1}^{N}\phi_i(y'_i)\prod_{(ij)\in\mathcal{E}}\phi_{ij}(y'_i, y'_j)$.

A specific assignment of values to Y is denoted by y and represented by the set $\{y_i^k\}$ of $K \cdot N$ indicator variables, for which $y_i^k = I(y_i = k)$. In order to foster the associativity of the model, we reward instantiations that have neighboring nodes, which are labeled by the same class. This is enforced by requiring $\phi_{ij}(k,l) = \lambda_{ij}^k$, where $\lambda_{ij}^k > 1$, for all $k = l$, and $\phi_{ij}(k,l) = 1$, otherwise [14]. Inference is carried out by solving the *maximum a-posterior (MAP)* inference problem, i.e., to find $\arg\max_{\mathrm{y}} P_\phi(\mathrm{y})$.

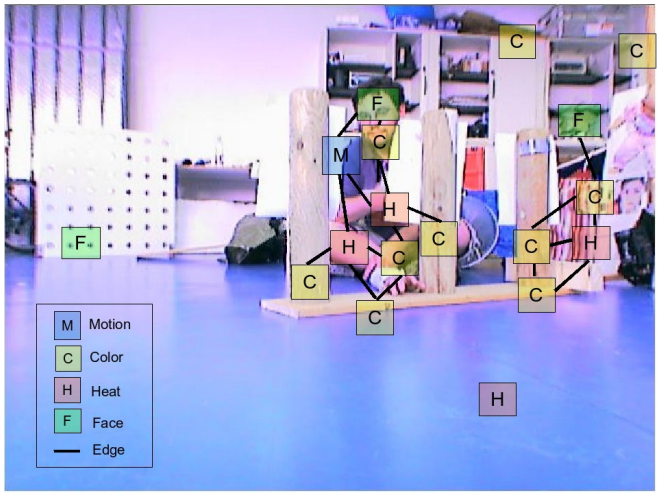


Fig. 2. MRF graph online constructed from features detected in the thermo and color images. Note that the number of shown features has been reduced for the sake of readability.

For our specific problem, we define the node potentials $\phi(y_i)$ by a vector of features which indicates the quality of the node's corresponding evidence frame. These are the size of the evidence frame, the number of included pixels, and the real world distance taken from the 3D range measurement. Likewise we define the edge potentials $\phi(y_i, y_j)$ by a vector of features that indicates the quality of neighborhood relations. Edges are built based on combinations of the evidence types introduced in Section II. In our case there exist 36 possible edge types, given the 6 different types of evidence. For example, the edge type *isWarmSkin* describes the combination of the features *heat* and *skinColor*. The feature vector of an edge includes the type of the edge, and the real-world distance measure between both nodes.

The MRF graph is dynamically constructed for each image from the video stream during runtime (see Figure 2 for an example). Firstly, we generate from the image data six sets of evidence frames, as described in Section II. From these sets, six types of MRF nodes are generated by calculating the feature vectors for each node potential, whereas each node type corresponds to one type of evidence. Secondly, edges between nodes are generated. Each node connects to the four closest neighbors in its vicinity, if they are within close real-world distance, which was set to a distance according to the size of human bodies. Finally, for each edge a feature vector for its edge potential is calculated.

For the sake of simplicity, we represent potentials by a log-linear combination $\log\phi_i(k) = \mathrm{w}_n^k \cdot \mathrm{x}_i$ and $\log\phi_{ij}(k,k) = \mathrm{w}_e^k \cdot \mathrm{x}_{ij}$, where x_i denotes the node feature vector, x_{ij} the edge feature vector, and w_n^k and w_e^k the row vectors according to the dimension of node features and edge features, respectively. Consequently, we can denote the MAP inference problem $\arg\max_{\mathrm{y}} P_\phi(\mathrm{y})$ by:

$$\arg\max_{\mathrm{y}} \sum_{i=1}^{N}\sum_{k=1}^{K}(\mathrm{w}_n^k \cdot \mathrm{x}_i)y_i^k + \sum_{(ij)\in\mathcal{E}}\sum_{k=1}^{K}(\mathrm{w}_e^k \cdot \mathrm{x}_{ij})y_i^k y_j^k. \quad (2)$$

Equation (2) can be solved as a linear optimization problem by replacing the quadratic term $y_i^k y_j^k$ with the variable y_{ij}^k and adding the linear constraints $y_{ij}^k \le y_i^k$ and $y_{ij}^k \le y_j^k$. Hence, the linear programming (LP) formulation of the inference problem can be written as:

$$\max \sum_{i=1}^{N}\sum_{k=1}^{K}(\mathrm{w}_n^k \cdot \mathrm{x}_i)y_i^k + \sum_{(ij)\in\mathcal{E}}\sum_{k=1}^{K}(\mathrm{w}_e^k \cdot \mathrm{x}_{ij})y_{ij}^k \qquad (3)$$

$$\text{s.t. } y_i^k \ge 0, \quad \forall i,k; \qquad \sum_k y_i^k = 1, \quad \forall i;$$

$$y_{ij}^k \le y_i^k, \quad y_{ij}^k \le y_j^k, \quad \forall ij \in \mathcal{E}, k,$$

which can, for example, be solved by the *Simplex* method. For a more detailed description, we refer to the work of Anguelov et al. [1]. Furthermore, it is necessary to learn the weight vectors for the node and edge potential from data, which we carried out by utilizing the *maximum margin* approach recommended by Taskar et al. [15].

IV. MODEL SELECTION

In the general case, solving the MAP inference problem, as shown in Section III, is NP-hard. Also in case of the considered two-class problem one notices an increase of computation time if the number of nodes and edges, and thus the size of the LP problem, grows.

Computation time is an important issue if applying the detection method, for example, to live video streams taken by a camera in a search and rescue scenario. Furthermore, the efficiency of specific evidence correlations (edge types in the MRF graph) depends on the scenario where the classifier is applied. For example, heat sources might be a stronger evidence for human bodies in an outdoor scenario as it would be in an indoor scenario with many heat sources, such as PCs and radiators. Therefore, our goal is to reduce the computation time of the MRF model by selecting edge types that significantly improve

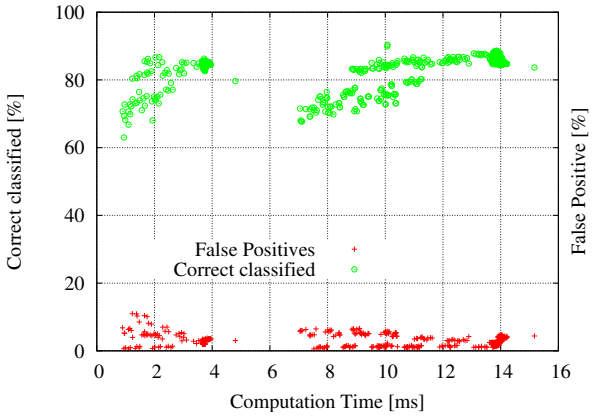


Fig. 3. Comparing model complexity (computation time) with the percentage of correctly classified vision features (green), and the percentage of false positives (red). Each data point corresponds to a MRF model with specific types of edges activated.

the classifier with respect to the data. For example, consider the situation in Figure 2. Since measurements from different sensors are generally more significant, the two edges between the motion node and the two heat nodes are more valuable as a single connection between both heat nodes only. However, by selecting the four closest nodes both heat nodes would be connected.

Therefore, we examined the contribution of specific edge types to the overall detection rate. This has been carried out by learning MRF models with different sets of activated edge types while measuring accuracy and computation time needed for inference. Figure 3 summarizes the result, where each data point corresponds to a specific combination of edge types. As can be seen, MRF inference needs between 2 ms and 16 ms, depending on the combination of activated edge features. Interestingly, a higher amount of computation time does not necessarily yield better classifier performance. Good classifier performance can already be achieved at a much smaller computation time than needed for computing models containing all types of edges, if the significant edge types are activated.

However, since the complexity of exhaustive search is in $O(2^n)$, and learning a single classifier takes a comparably high amount of time, finding optimal edge types is intractable in the general case. Therefore, we developed a genetic algorithm for selecting most efficient combinations of edges. The scoring function for guiding the search has been defined by the trade-off between classifier performance and computation time:

$$S = U - \alpha C(p_i), \tag{4}$$

where U corresponds to the utility metric, $C(.)$ denotes a cost function reflecting the model complexity, p_i denotes the ith permutation, and α is a parameter regulating the trade-off. Depending on the application domain, U can be computed, for example, from the negative *false-positive* rate, the total number of correctly classified evidence frames, or the percentage of the correctly classified area. Without loss of generality, we decided to use the area-based utility metric since it enforces the detection of body silhouettes rather than frames on their own. The series in Figure 4 (c) depicts this metric by the blue clusters for true positives, which have been build by the union of all true positive evidence frames in the respective image.

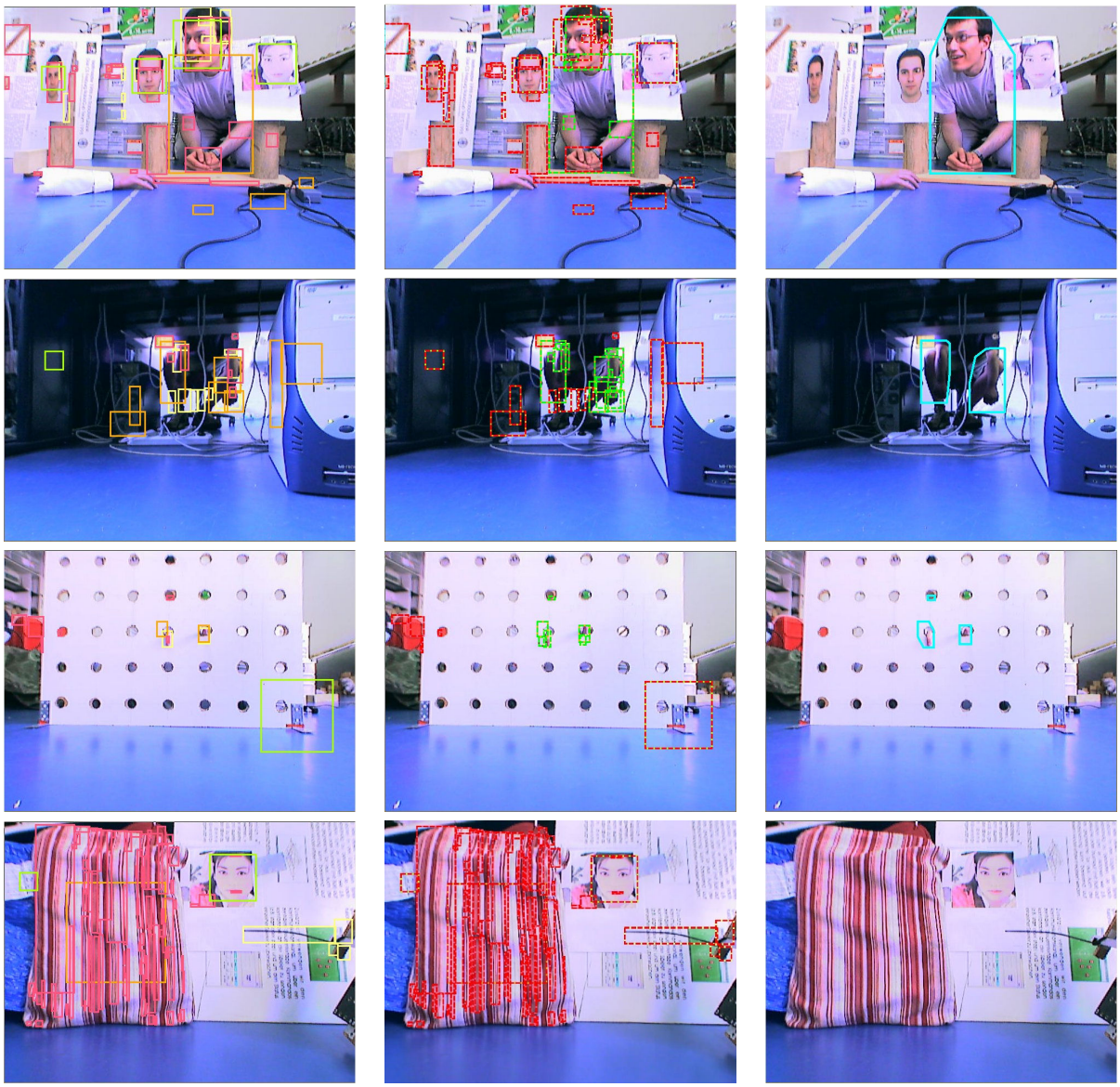
The scoring function is utilized as fitness function for the genetic algorithm (GA). Solutions, i.e, specific combinations of edge types, are represented for the genetic optimization as a binary string. Each edge type is represented by a bit, and set to *true* or *false* regarding the activation of the corresponding edge type. In order to guarantee that good solutions are preserved within the genetic pool, the so-called *elitism* mechanism, which forces the permanent existence of the best found solution in the pool, has been used. Furthermore, we utilized a simple one-point-crossover strategy, a uniform mutation probability of $p \approx 1/n$, and a population size of 10. In order to avoid that solutions are calculated twice, all computed solutions are memorized by their binary string in a hash map.

V. EXPERIMENTS

We generated more than 6000 labeled examples from video streams recorded within a NIST arena-like environment and split them into three folds. The training data contains true evidence, which is exclusively generated from human bodies, and false evidence, which is generated from artificial sources, such as a heat blanket, laptop power supply, printouts of faces, moving objects, and objects with skin-like texture, such as wood. Figure 4(a) depicts some examples from the training data[2]. Each color frame corresponds to an evidence type. Green frames correspond to face detection, orange frames to heat detection, red frames to color detection, and yellow frames to motion detection. Note that the training data contains intentionally many cases in which the vision system produces ambiguous evidences. From this data, MRF models were trained by K-fold cross-validation, with $K = 3$.

In order to evaluate the model selection, we reduced the total set of edge types from 32 to 9 since in our case, some feature combinations represent the same concept, as for example the combinations of *heat* from the thermo images together with *face* from the color images, and *face* and *heat* both from the thermo image. The genetic model selection has been evaluated by multiple runs with varied parameter α and varied scoring metric (Equation (4)), e.g., based on the false-positive rate, total error, and total area. In average, the genetic selection yielded the optimal solution after considering 60 ± 7 models, which is more than eight times faster than performing exhaustive search over all 512 possible models.

[2]The according video can be found under `http://www.informatik.uni-freiburg.de/~kleiner/video/iros_vision.mpg`



(a) (b) (c)

Figure 4. Examples from the test data: (a) Detected evidences: skin color (red), heat (orange), motion (yellow), face (green), (b) the same evidences after classification and (c) all positive evidences clustered into areas.

For the selection of the final MRF model, we utilized the area-based scoring metric with $\alpha = 2.0$. The genetic algorithm selected a classifier which activates, for example, the edge types $motion \wedge face$, $heat \wedge skin$, $heat \wedge face$, $heat \wedge motion$, and forbids $motion \wedge skin$, as well as all edge types between the same kind of nodes. Finally, the selected classifier reached an accuracy of 87.9% at 2.3 ms, in contrast to the classifier with all edges activated (88.8% at 12.6 ms) and the classifier with no edges activated (71.3% at 1.1 ms). Figure 4 shows some examples of the successful application of the classifier even to hard cases, such as small finger movements, and test persons completely surrounded by faked evidences.

We compared the performance of the optimized MRF model with a Support Vector Machine (SVM) [16] based classifier. The SVM has been trained on the same features as they were generated for the MRF model, shown in Section II. In Table I the performance for classifying single evidence frames of both classifiers is reported. The

	SVM			MRF		
	False Pos.	False Neg.	Err. [%]	False Pos.	False Neg.	Err. [%]
Human	23	433	39.4	26	143	14.6
Faked	758	0	11.3	151	0	2.3
Both	703	2689	32.8	484	836	12.8
Total	1484	3122	21.0	661	979	7.5

Table I. Comparison of the SVM and MRF classifier: Numbers denote the amount of wrongly classified evidences in images containing humans, faked evidence, and both

results have been partitioned into three sets showing the performance on examples containing human evidence, faked evidence, and both. The result indicates that the optimized MRF model performs better in terms of false-positive classifications, particularly in situations containing exclusively faked data.

In the context of search and rescue it is desirable to reach a high true-positive rate on each image, i.e., humans are detected reliably, and a low false-positive rate, i.e., no victim alarm from faked evidence. This is not directly expressed by the percentage of correctly classified frames since one wrongly detected frame within an image suffices to trigger the false alarm. Therefore, we counted the true-positive and false-positive rate for both classifiers *image-wise*, i.e., images are counted as true-positive, and false-positive, if there is a single correct and a single wrong evidence found, respectively. It turned out that for images containing human evidence, both MRF and SVM reported a victim correctly in 100% of the cases, whereas in images containing faked evidence, the SVM wrongly reported a victim in 60% and the MRF in 13% of the cases. Note that the result of the MRF model is comparably good since the training data also contained images with more than three different types of faked evidence at the same time, which makes a distinction from human beings impossible.

VI. CONCLUSION

We introduced a system that creates MRF models in real-time from motion, color, and shape evidence, detected by a CCD camera and IR camera, respectively. In order to reduce computational demands during inference, the building process of models has been optimized by a genetic algorithm, which decides offline relevant edge types with respect to the data. Finally, the selected classifier was five times faster than the model with all edge types activated, while gaining optimal performance in terms of the complexity trade-off, and near-optimal performance in terms of accuracy. We compared the optimized model with a SVM and showed that the false-positive rate has been significantly reduced, which is an important aspect when considering victim detection in the context of rescue robotics. From an image-wise evaluation of the classifier it can be concluded that the approach reliably detects victims if present and only in hard cases, i.e., if the number of faked evidences is high, false-positives occur. The classifier performance could be further improved by introducing temporal relations, i.e., by adding edges between evidences found in preceding images from the video stream.

The proposed approach is general as it can easily be extended for incorporating other types of human evidence, such as audio noise, e.g., tapping, and CO_2 emission. Also given these evidence types, it might be interesting to figure out which correlations significantly contribute to the classifier's performance. In future work, we will furthermore consider to extend the class variable by classes describing the victim's state, e.g., *aware* and *unconscious*, which can generally be concluded from correlations between evidence types.

References

[1] D. Anguelov, B. Taskar, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz, and A.Y. Ng. Discriminative learning of markov random fields for segmentation of 3D range data. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, California, June 2005.

[2] S. Bahadori and L. Iocchi. Human body detection in the robocup rescue scenario. In *CDROM Proc. Int. RoboCup Symposium '03*, 2003.

[3] C. Beleznai, B. Frühstück, and H. Bischof. Human detection in groups using a fast mean shift procedure. In *Proc. IEEE Int. Conf. on Image Processing*, pages 349–352, 2004.

[4] Gary Bradski. The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 25(11):120, 122–125, November 2000.

[5] James Bruce. Realtime machine vision perception and prediction. Carnegie Mellon University, Undergraduate Thesis, 2000.

[6] D. Cao, O. Masoud, D. Boley, and N. Papanikolopoulos. Online motion classification using support vector machines. In *IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 2291–2296, 2004.

[7] R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):781–796, 2000.

[8] S.C. Dass, A.K. Jain, and X. Lu. Face detection and synthesis using markov random field models. In *Proc. of the 16th Int. Conf. on Pattern Recognition (ICPR'02)*, volume 4, pages 201–204, Washington, DC, USA, 2002. IEEE Computer Society.

[9] J. Han and B. Bhanu. Detecting moving humans using color and infrared video. In *Proc. IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI 2003)*, pages 228–233, 2003.

[10] A. Jacoff, E. Messina, and J. Evans. Experiences in deploying test arenas for autonomous mobile robots. In *Proceedings of the 2001 Performance Metrics for Intelligent Systems (PerMIS) Workshop*, 2001.

[11] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. *IEEE ICIP*, 1:900–903, September 2002.

[12] I. Nourbakhsh, M. Lewis, K. Sycara, M. Koes, M. Yong, and S. Burion. Human-robot teaming for search and rescue. *IEEE Pervasive Computing*, 4(1):72–78, January 2005.

[13] H. Sidenbladh. Detecting human motion with support vector machines. In *Proc. of the 17th Int, Conf. on Pattern Recognition*, volume 2, pages 188–191, 2004.

[14] B. Taskar, V. Chatalbashev, and D. Koller. Learning associative markov networks. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)*, 2004.

[15] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *Advances in Neural Information Processing Systems (NIPS 2003)*, Vancouver, Canada, 2004.

[16] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[17] P. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE CVPR*, 2001.

[18] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[19] Y. Wu and T. Yu. A field model for human detection and tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5), 2006.