2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Daejeon Convention Center
October 9-14, 2016, Daejeon, Korea

Task-conversions for Integrating Human and Machine Perception in a Unified Task

Hyungtae Lee[1,2], Heesung Kwon[2], Ryan M. Robinson[2], Daniel Donavanik[2],
William D. Nothwang[2], and Amar R. Marathe[3]

***Abstract*— The different strategies for feature extraction and synthesis employed by humans and computers are often complementary, hence combining the two into an integrated object recognition system may considerably improve performance over either used in isolation. Rapid Serial Visual Presentation (RSVP) is one well-established technique that has shown promise integrating human perception into a machine perception system. In this paper, we apply computer vision techniques to image data filtered through human RSVP. We introduce "task conversions" to integrate the two modalities, applying the precise localization capabilities of computer vision with the detection capabilities of RSVP. We employ naive Bayesian fusion and a novel method, dynamic belief fusion (DBF), in a joint scheme as fusion approaches. Preliminary experiments demonstrate that DBF extracts complementary information from both human and machine sources to improve performance for both target classification and object detection.**

I. Introduction

Humans are unparalleled in their ability to recognize objects against complex or cluttered backgrounds. However, human perception is limited in throughput, and may be substantially impacted by factors such as fatigue, boredom, and heavy cognitive workload. Furthermore, attempts to exploit human processing directly through the use of neurophysiological signals suffer a range of challenges which in most cases render them inferior to near real-time GPU implementations of computer vision algorithms [1].

In this paper, we demonstrate that fusion of human detection with computer vision can enhance detection performance. We make use of the Rapid Serial Visual Presentation (RSVP) paradigm used by Touryan et al. [2]. In RSVP experiments, participants are instructed to press a button when target images are seen among images presented at a rate of 5Hz. EEG signals are concurrently monitored for a signature which occurs after presentation of a target image, indicating a positive detection. The responses to RSVP denote the presence or absence of a target, but do not identify the specific target type or localize it within an image. Computer vision may be used to provide this additional information by applying specific target object models against location hypotheses in the image.

[1]Booz Allen Hamilton Inc., McLean, VA, USA lee_hyungtae@bah.com
[2]Sensors & Electron Devices Directorate, Army Research Laboratory, Adelphi, MD, USA {heesung.kwon.civ, ryan.robinson14.ctr, daniel.donavanik.ctr, william.d.nothwang.civ}@mail.mil
[3]Human Research & Engineering Directorate, Army Research Laboratory, Aberdeen Proving Ground, MD, USA amar.marathe.civ@mail.mil

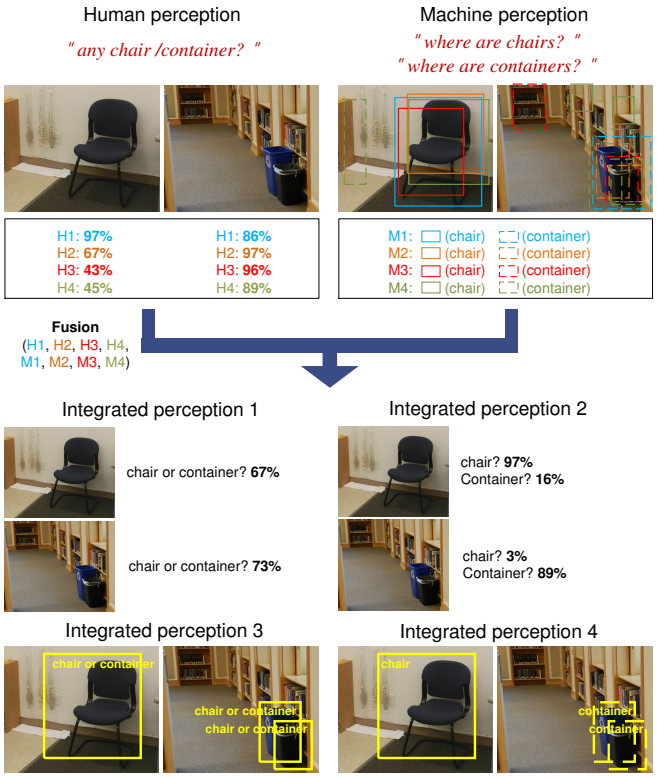

Fig. 1. **Integration of human and machine perception in four different tasks:** 4 approaches using human perception (H 1∼4) and 4 approaches using machine perception (M 1∼4). *Top left:* query is given to the human subject. *Top right:* machine perception results superimposed on image. *Bottom:* fusion outputs of integrated human and machine perception.

Using the unique conditions and capabilities of the two modalities, a family of algorithms was created to perform four related but distinct "tasks": (i) presence or absence of any target in an image (ii) presence or absence of a specific target type within an image, (iii) presence or absence and location of a target within an image, and (iv) presence or absence and location of a specific target type in an image. Figure 1 demonstrates queries given to the human and machine, responses, and fusion results for each of the four tasks. Due to the binary (presence or absence)-only nature of the RSVP paradigm, the human is only directly able to perform Task 1[1], while the computer can hypothetically perform all four tasks. One way to mitigate this shortcoming

[1]For this RSVP study, we treat incomplete results from Task 2 as complete results for Task 1. This has the effect of artificially deflating human performance figures, but does not impact the validity of the fusion result. See discussion in Experiments and Conclusion.

978-1-5090-3762-9/16/$31.00 ©2016 IEEE

of RSVP is to convert the information provided to a new form in order to encode the necessary information in an RSVP-compatible task. To that end, we introduce "task-conversion" strategies to allow for a meaningful human response in any of the above four tasks, and apply a combination of fusion approaches to exploit these jointly with computer vision detections.

The rest of this paper is organized as follows. The proposed task-conversion and fusion strategies are described in Section II. Section III provides the detail of human decision classifiers and computer-vision-based object detectors selected for fusion. The proposed partition of the office object dataset and experimental results are presented in Section IV.

A. Related Work

The Rapid Serial Visual Presentation (RSVP) paradigm traces its origins to the work of Potter and Levy [5], who originally developed the approach to test the amount of information that human subjects could absorb at speeds that would not allow for deep periods of cognitive fixation. Early research was heavily biased towards text processing [6], with a concurrent shift in latter decades towards practical applications in human-in-the-loop information processing. Mills and Weldon [7] proposed an RSVP-driven framework for dynamic text presentation which demonstrated mixed results versus other speed-reading approaches.

In the past two decades, there has been an accelerating body of research in the use of RSVP for human-in-the-loop image processing. DARPA's Neurotechnology for Intelligence Analysts program (NIA)[10] applied the technique, along with EEG processing, to the reduction in the search space of large amounts of previously unindexed imagery by human analysts [8]. Fei-Fei et al. [11] made the distinction between identification of *targets* in a scene, generally well-correlated with the P300 EEG signal (onset at 300ms after stimulus presentation), and the *gist* of images, which could be reliably detected in RSVP after a presentation of less than 100ms. Evans and Treisman [12] found that when the contrast between foreground and background objects is strong, i.e. man-made objects with regular geometry against a natural background, average time to detection in RSVP was as little as 113ms. Other groups [2], [8], [9] have demonstrated broad success in the use of RSVP to increase the throughput of human subjects analyzing imagery.

Recently, groups have attempted to integrate human-in-the-loop processing in a fully closed-loop system. Branson et al. [13] propose a "twenty questions" paradigm in which positive human response to one or more images in an RSVP set flags those images for downstream processing as a means of disambiguating between closely related classes of targets. Other attempts [14], [15] at fused human-machine recognition systems generally interleave human interaction with an "active learning" phase, sequentially asking users to label examples in order to steer the underlying algorithm. Our previous work integrates human classifiers with computer-vision-based detectors by a novel DBF approach in a heuristic way in object detection [16]. The task-conversion we addressed in this paper allows fusion in all four tasks.

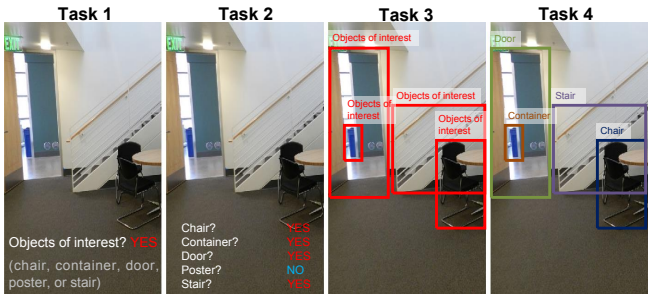

Fig. 2. **Examples of four tasks:** Task 1, 2, 3, and 4 are categorizing images as (1) containing any of the objects-of-interest without localization, (2) classifying images for each object-of-interest class without localization, (3) categorizing images as containing any of the objects-of-interest with localization, and (4) classifying images from each object-of-interest class with localization, respectively.

II. Task Conversion and Fusion Strategies

A. Task-conversion techniques

The objective of this study is to effectively fuse responses from human and machine vision approaches. We identify four tasks that may be performed given human and machine responses to images, illustrated in Figure 2:

1) detect images containing any of the objects of interest,
2) detect images containing a specific object of interest (for each object category),
3) detect location of any of the objects of interest,
4) detect location of a specific object of interest (for each object category).

As previously stated, there is an inherent challenge in that human responses in the above experimental design do not yield identity or location information for specific targets. The human is therefore only directly able to perform Task 1, while the computer can hypothetically perform all four tasks. In operation, object detection algorithms generate candidate windows of various size and location within the image, and search for target features within each window; the output is a scored set of windows performed; when this procedure is done for each specific target class, it is analogous to Task 4.

Figure 3 demonstrates task conversion strategies for employing the output of computer-vision-based object detectors in all four tasks. Initially, detectors search possible windows in the image for a particular object and assign each a detection score, i.e. Task 4. Each object of interest has an associated detector and the confidence scores of the detectors can not be compared directly due to *difference in scale*, i.e. score range. We employ Platt scaling [17] which calibrates the results of multiple detectors in order to allow them to be compared. Platt scaling is the process of rescaling and shifting the decision boundary of classifiers to create one unique boundary. [18] demonstrates that Platt scaling greatly improves the inter-detector ordering while each decision boundary is no longer an optimal solution for learning each detector. Platt scaling learns parameters α and β, which

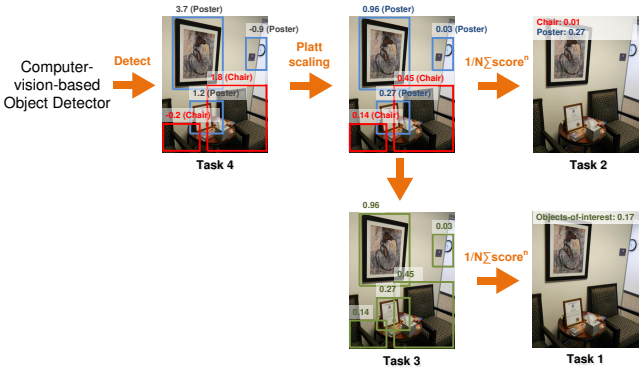

Fig. 3. Task conversion for Computer-vision-based object detectors

Fig. 4. Task conversion for human perception and precision and recall curve for all four tasks. Precision and recall are calculated for Subject 1's peception ability. For Task 2 and 4, the precision-recall curve for the "chair" category is shown.

are used in fitting a probability distribution of outputs of detectors to a shared validation set. (Here, the probability distribution is assumed to follow a sigmoid function with two parameters α and β.) The calibrated score f_s for the detection with detection score sc of the i^{th} detector is as follows:

$$f_{s,(\alpha_i,\beta_i)}(sc) = \frac{1}{1+e^{\alpha_i sc+\beta_i}} \tag{1}$$

Details of implementation are described in [17]. Note that this generates a confidence score between 0 and 1 for each detector's candidate window(s) (as in the top-middle image in Figure 3) and preserves these scores when converting to a class-agnostic representation (Task 3) (as in the bottom-middle image in Figure 3).

We use the following aggregation formula to convert detection-level scores from Platt-scaled Task 4 (or Task 3) to a single image-level score in Task 2 (or Task 1):

$$score_I(H) = \frac{1}{N}\sum_{j=1}^{N} score_j(H)^k \tag{2}$$

where $score_j(H)$ is the score of j^{th} detection for a certain hypothesis H localized on image I, N is the number of detections in the image, and $k \geq 1$ is an empirical parameter. We choose a constant N for each image (15 in evaluation), resulting in equal re-weighting across all images. If the original number of detections in an image is larger than N, only the N top-scoring detections are selected, and if the number is less than N, we consider the missing detections to be zero-scoring. We use the value of $k=5$, which empirically proved to be effective in converting from a detection task to a classification task in [19]. Note that a higher k increases the contribution of high-scoring detections compared to lower-scoring detections. This aggregation formula is used for converting from scaled confidence for each object to Task 2 as well as from Task 3 to Task 1.

Figure 4 demonstrates task conversion strategies for the human response. It is impossible to directly infer object identity or location from the output of classifiers because identification and localization are more difficult than pure classification. Similarly, estimating the location of objects of interest from the output of classifiers is also impossible. The performance of tasks requiring more detailed inference should be worse, as in the sample PR curves shown in a right-most image of Figure 4. (In a PR curve, greater area under the curve (AUC) denotes higher precision.) In a training step, we are able to identify classifiers and detectors which perform poorly in comparison to others. In a later subsection, we will introduce a fusion approach referred as *dynamic belief fusion* (DBF) that effectively integrates human decisions with computer-vision-based object detectors in all four tasks by treating the portion of performance caused by these factors as "uncertainty".

B. Integrating outputs of multiple classifiers/detectors

Figure 5 illustrates the strategy used to fuse multiple classifiers used in Tasks 1 and 2. In Tasks 1 and 2, each classifier produces only one score per image, corresponding to the target hypothesis. To integrate the outputs of multiple classifiers, we fuse the image-level scores $\mathbf{s} = [s_1\ s_2\ \cdots\ s_K]$ where s_i is the output score of the i^{th} detector. K is a number of classifiers.

Figure 6 illustrates how detection-level scores (as produced in Tasks 3 and 4) from multiple detectors are fused. In Tasks 3 and 4, each detector outputs and scores multiple candidate windows per image for the target hypothesis. We cluster detection windows which possibly contain the same target at the same general location and generate the score vector $\mathbf{s} = [s_{1j_1}\ s_{2j_2}\ \cdots\ s_{Kj_K}]$ for each cluster, where s_{ij_i} is the output score of j_i^{th} detection window of an i^{th} detector. Here, we consider two windows to be placed in the same location if the intersection over the union of their bounding boxes is over 0.5. Since clustering is performed for all individual windows of multiple detectors, multiple clusters corresponding to the same target hypothesis can exist. (For example, three clusters, each of which is detected in each image, contain the same chair in Figure 6.) If a cluster contains more than one detection from the same detector, only the maximum score is inserted to the corresponding bin of the score vector. If a particular detector does not contain an overlapping detection window (of a particular target class) where other do, $-\infty$ is inserted to the corresponding bin, indicating no detection information is provided to the fusion from the detector. After calculating fusion score for all window clusters, we employ non-maximum suppression to remove the clusters with lower fusion scores than other clusters in the same location.

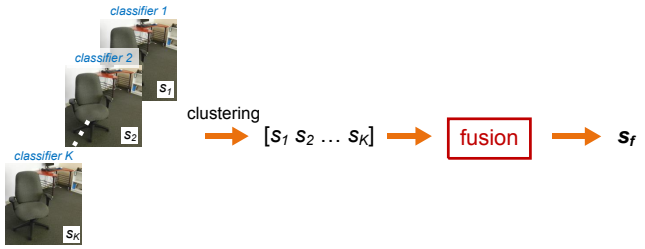

Fig. 5. Clustering process for Task 1 and 2.

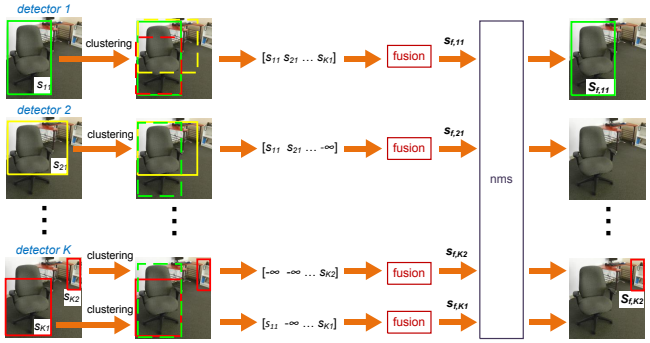

Fig. 6. Clustering process for Task 3 and 4.

C. Fusion approaches: Naive Bayesian Fusion and Dynamic Belief Fusion

We employ two probabilistic approaches, naive Bayesian fusion [3] and DBF [4], to create the final score vector $\mathbf{s}$.

Naive Bayesian Fusion: naive Bayesian fusion assumes the classification and/or detection approaches to be fused are independent. The combination law, known as Bayes Rule, is given as:

$$p(c|\mathbf{s}) \propto p(c)p(\mathbf{s}|c), \tag{3}$$

where c is a particular hypothesis, $\mathbf{s} = [s_1, s_2, ..., s_K]$ is the set of recognition scores from the constituent classifiers/detectors, and $p(c)$ and $p(\mathbf{s}|c)$ are the prior probability of hypothesis c and the likelihood of score $\mathbf{s}$ given the hypothesis, respectively. In this case, the possible hypotheses are presence and absence. These hypotheses can be applied at the image level (as obtained from the human RSVP response for any object category) or at the detection level (as obtained from machine perception for specific object categories). With the assumption of independence, a joint likelihood can be developed as the product of the likelihoods of K approaches:

$$p(\mathbf{s}|c) = \prod_{i=1}^{K} p(s_i|c). \tag{4}$$

The opposite hypothesis, $p(\neg c|\mathbf{s})$ can be calculated in the same manner. A model containing $p(c)$, $p(\neg c)$, $p(s_i|c)$ and $p(s_i|\neg c)$ was generated by aggregating scores in a validation set. During testing, prior and likelihood information were determined for each approach by referencing the model, and the final "fused" score was calculated as $p(c|\mathbf{s}) - p(\neg c|\mathbf{s})$.

Dynamic Belief Fusion (DBF): DBF [4] is an approach we previously developed to assign probability to hypotheses dynamically under the framework of Dempster-Shafer Theory (DST) [20], [21]. DST is based on Shafer's belief theory [21]. Considering the two hypotheses, target (c) and non-target ($\neg c$), it assigns probabilities that directly support those hypotheses, and also instantiates an intermediate state I which represents evidence that could plausibly support either hypothesis. This intermediate state is given its own probability, quantifying the level of ambiguity that makes either hypothesis plausible. The belief function $bel(A)$ for a set A can be defined as

$$bel(A) = \sum_{B|B\subseteq A} p(B). \tag{5}$$

If the probability is assigned to each of the three hypotheses, $bel(c) = p(c)$ and $bel(\neg c) = p(\neg c)$. (Note that $bel(I) = p(c) + p(\neg c) + p(I)$.) Once all classification and detection approaches assign probability to each hypothesis (including the intermediate state), Dempster's combination rule [20] can be applied to calculate a joint probability:

$$p_1 \oplus p_2(c|s_1,s_2) = \frac{1}{L} \sum_{X\cap Y=c, c\neq\emptyset} p_1(X|s_1)p_2(Y|s_2), \tag{6}$$

where $L = \sum_{X\cap Y\neq\emptyset} p_1(X|s_1)p_2(Y|s_2)$. L is the sum total of probability mass whose common evidence is not the null set. Dempster's rule can be extended for multiple approaches using the associative and commutative properties of probabilities (i.e. $p_f = p_1 \oplus p_2 \oplus \cdots \oplus p_K$) with the following formula:

$$p_f(c|\mathbf{s}) = \frac{1}{L} \sum_{X_1\cap X_2\cap\cdots\cap X_K=c, c\neq\emptyset} \prod_{i=1}^{K} p_i(X_i|s_i), \tag{7}$$

where $L = \sum_{X_1\cap\cdots\cap X_K\neq\emptyset} \prod_{i=1}^{K} p_i(X_i|s_i)$.

For the i^{th} classifier/detector, probabilities for a set of hypotheses $\{c, \neg c, I\}$ are calculated using precision and recall information calculated in a validation set. For a given score s, the corresponding probabilities are given by:

$$\begin{aligned} p_i(c|s) &= prec_i(s), \\ p_i(\neg c|s) &= 1 - prec_{bpd}(s) = rec_i(s)^n, \\ p_i(I|s) &= prec_{bpd}(s) - prec_i(s) \\ &= 1 - rec_i(s)^n - prec_i(s), \end{aligned} \tag{8}$$

where $prec_i$ and rec_i are precision and recall for the i^{th} approach, respectively. $prec_{bpd}$ is the precision of a theoretical best possible detector which is assumed to have no ambiguity, and is defined as $1 - rec_i(s)^n$, where n is a parameter obtained by cross-validation and shared between all classifiers/detectors. [4] demonstrated the effectiveness of the best possible detector for fusion. After $prec_i$ and rec_i are computed in a validation set, testing is performed. Values corresponding to the test recognition score are used in the individual probability assignments for $\{c, \neg c, I\}$. Similar to naive Bayesian fusion, $p_f(c|\mathbf{s}) - p_f(\neg c|\mathbf{s})$ is used as the final "fused" score.

III. Approaches Analyzing Human Perception and Machine Perception

In the proposed fusion approach, we employ three neural classifiers, button press, and four computer vision object detectors.

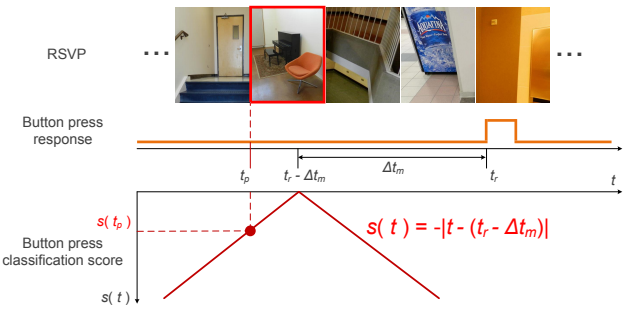

Fig. 7. **Button press classification score computation:** the first and second row demonstrate images presented by RSVP and button press response of participant when looking at a target, respectively. Button press classification score computation is shown in the third row.

A. Approaches analyzing human perception

Human visual perception can be estimated via biometric signals such as EEG or button press. For EEG, we employ three standard neural classification algorithms, listed below:

- Hierarchical Discriminant Component Analysis (HDCA) [22]
- xDAWN+Bayesian Linear Discriminant Analysis (XD+BLDA) [23], [24]
- Common Spatial Patterns with Bayesian Linear Discriminant Analysis (CSP+BLDA) [23], [24]

Each classifier is described in Marathe et al. [25] These classifiers require a training step in which actual responses are paired with ground-truth labels (target vs. non-target), and the feature space dominated by target responses is separated from the space dominated by non-target responses.

Compared with noisy EEG, button presses are a more concrete indication of a classification decision. However, in the case that images are presented at a rapid rate, as is generally the case in RSVP, the timing of the button press relative to the stimulus presentation is usually delayed and irregular [26]. The likelihood that an image resulted in a button press is therefore some function of the time delay between presentation and button press. The button press classification score $s(t_p)$ at an image presentation time t_p is calculated by

$$s(t_p) = -|t_p - (t_r - \Delta t_m)|, \tag{9}$$

where t_r and Δt_m is the button press response time and the median reaction time, respectively. Δt_m is empirically calculated from the training set. Figure 7 shows the button press classification score computation.

B. Machine learning approaches for machine perception

Four computer-vision-based object detectors were selected for fusion with human decision, and are summarized below:

- Histogram of Oriented Gradients + Support Vector Machine (HOG+SVM) [27]: HOG features are employed to represent object appearance in terms of a distribution of gradients. An SVM is then trained to distinguish object from background.
- Exemplar SVMs (ESVM) [18]: ESVM learns a SVM-based separate classifier for each positive training image (called as an exemplar) using a HOG feature, and scores candidate detections based on "distance" to exemplars.
- Deformable Part Models (DPM) [28]: Objects are represented as sets of parts that can be deformed using HOG features at two scales and latent features, with a deformation cost.
- Fine Tuned Convolutional Neural Network (FT-CNN) [19]: FT-CNN is based on AlexNet [29] pre-trained on a very large image dataset, ImageNet [30]. The target image dataset used in this work contains much fewer images than ImageNet with quite different visual characteristics. To adapt the CNN structure of AlexNet to category distribution and characteristics of the target dataset, the final fully connected layer referred a classification layer is learned again over the target dataset.

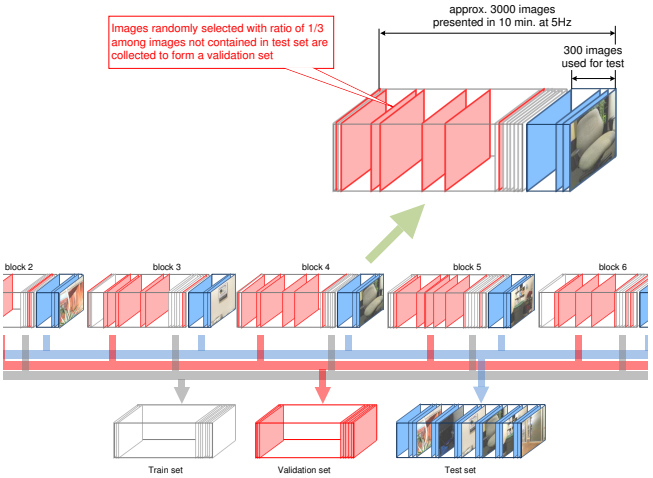

Fig. 8. **The proposed partition of the image set used in the RSVP task:** the image set consists of 6 blocks. For each block, the last 300 images are used for testing. The images not contained in the test set are randomly split into training and validation sets at a 2:1 ratio.

Each of these techniques uses distinct principles for feature extraction and synthesis; we expect this will lead to the generation and fusion of complementary information.

IV. Experiments

A. RSVP dataset and data partition for evaluation

The RSVP experiment used in this analysis presented images at 5 Hz (200 ms per image). The complete experiment consisted of 6 blocks of 10 minutes each (approximately 3000 images were selected for each block). Images in each block were randomly chosen, but contained a specific ratio of target/non-target images (this was a variable of interest in the preceding study). Six different ratios (0.01, 0.03, 0.05, 0.07, 0.09, and 0.11) were randomly assigned to the six blocks. Target images depicted at least one of five object categories: chair, container, door, poster, and stair. Due to the relatively small size of the dataset, images (mostly non-target images) were repeated within blocks and between blocks.

Fifteen subjects participated in the RSVP experiments. The subjects were instructed to watch the sequence of images and

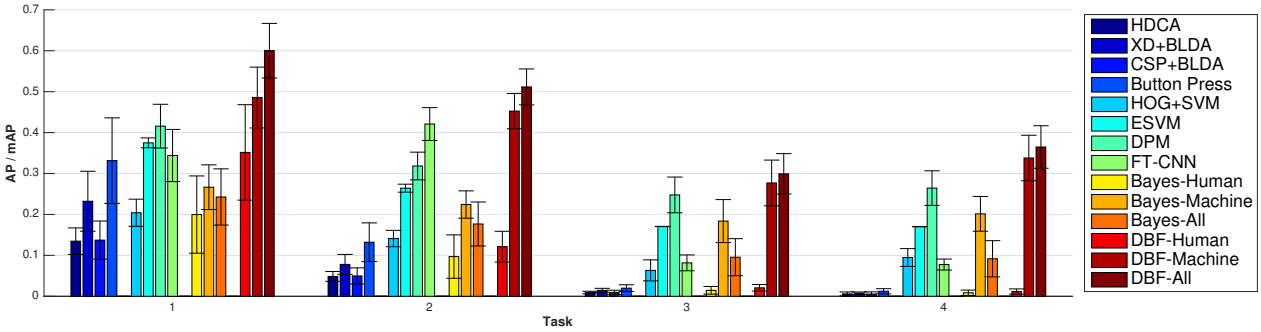

Fig. 9. **Performance comparison of individual approaches and fusion approaches (Bayesian fusion and DBF):** for each fusion approach, three bars indicating results integrating of human perception approaches only, computer-vision-based apporaches only, and all perception approaches. Fusion is performed in four tasks and their results are shown in order. Error bars denote the standard deviation across subjects.

to press a button when an object of interest was seen, for each target category (Task 2); due to incomplete performance of the task through the range of object categories, the results were ultimately treated as equivalent to Task 1. EEG data were collected in parallel using a BioSemi Active Two system with 256 channels (downselected to a subset of 64 channels that most closely matched electrode locations in the standard 10-10 arrangement), digitally sampled at 1024 Hz. Offline, the EEG data were decimated to 256 Hz and digitally band-pass filtered between 0.5 and 50 Hz. The neural and button press classifiers were trained and tested through a cross validation which preserved independence of these two sets in each cross validation.

In this experiment, partitioning of images for the RSVP task was random, with a fixed ratio of targets to distractors for each subject to enable the demonstration of the human-computer-vision fusion concept. In subsequent work, we will further explore the effect of target to distractor ratios as well as individual choices of images by repeating multiple randomizations across subjects.

Unlike human perception, in which the response to an image can be altered by the subject's physiological state (fatigue, workload, etc.) or influenced by the preceding images, a (non-adaptive) computer-vision-based algorithm will consistently produce the same outputs for the same image. As previously stated, computer vision algorithms cannot be trained, validated, or tested using repeated images due to the possibility of overfitting. Thus, the training, validation, and test sets were separate and non-overlapping.

A specific dataset partitioning procedure was performed to generate common validation and test sets, illustrated in Figure 8. The final 300 images from each RSVP test block were used as the fusion test set, as we hypothesize that the data at the end of each RSVP block elicit a steady-state subject performance level (to be validated in future work). The remaining RSVP images were collected to form a training set used for training the computer vision object detectors and a validation set used for computing the prior performance and likelihood model for each approach: essentially "fusion training". The partitioning between training and validation sets was done randomly at a 2:1 ratio. If repeat images from the RSVP task are not counted, the ratio among the train/validation/test sets was 2:1:2.

B. Performance comparisons

Performance was evaluated using average precision (AP) for Tasks 1 and 3, and using mean average precision (mAP) across object categories for Tasks 2 and 4. AP is a standard metric in the computer vision field, obtained by averaging precision values across the range of recall; in that sense, it is semantically similar to the AUC metric. AP/mAP results from naive Bayesian fusion and DBF were calculated and averaged over all subjects. The fusion performance of human-only, machine-only, and human+machine (i.e. "all") perception are shown in Figure 9. For all tasks, we employed one-way ANOVA tests to assess the effect of fusion method on AP/mAP. Each one-way ANOVA considered the choice of individual approaches (e.g., Deformable Parts Model) or fusion approaches (e.g. DBF-Machine) as a main effect (14 approaches total). The results across 15 subjects are as follows: $[F_{(13,182)} = 73.87,\ p < 0.001]$, $[F_{(13,182)} = 375.27,\ p < 0.001]$, $[F_{(13,182)} = 237.39,\ p < 0.001]$, and $[F_{(13,182)} = 378.41,\ p < 0.001]$ for Tasks 1, 2, 3, and 4, respectively. These results imply that, at the very least, there were statistically significant differences between the best and worst approaches in each task. As a follow-on test, we compared the statistical difference between any pair of 14 approaches by using a multiple comparisons test. Interestingly, although the statistical difference between DBF-Machine and the best individual approach (DPM or FT-CNN, depending on the task) is not significant in Tasks 1, 2, and 3, by fusing human and machine perception (DBF-All), performance *does* yield statistically greater performance than all other approaches. Note that this occurs even in Tasks 3 and 4, where human-only performance with DBF is very poor. These results support our hypothesis that human and machine perception yield complementary information that can be leveraged for improved performance. Note that Bayesian fusion is statistically lower-performing than the best individual approach in all four tasks. Bayesian fusion is negatively influenced by poor-performing approaches (i.e. human perception) and cannot adequately resolve conflicting information. Furthermore, the performance of human+machine Bayesian fusion consistently drops below machine-only fusion; the opposite is true

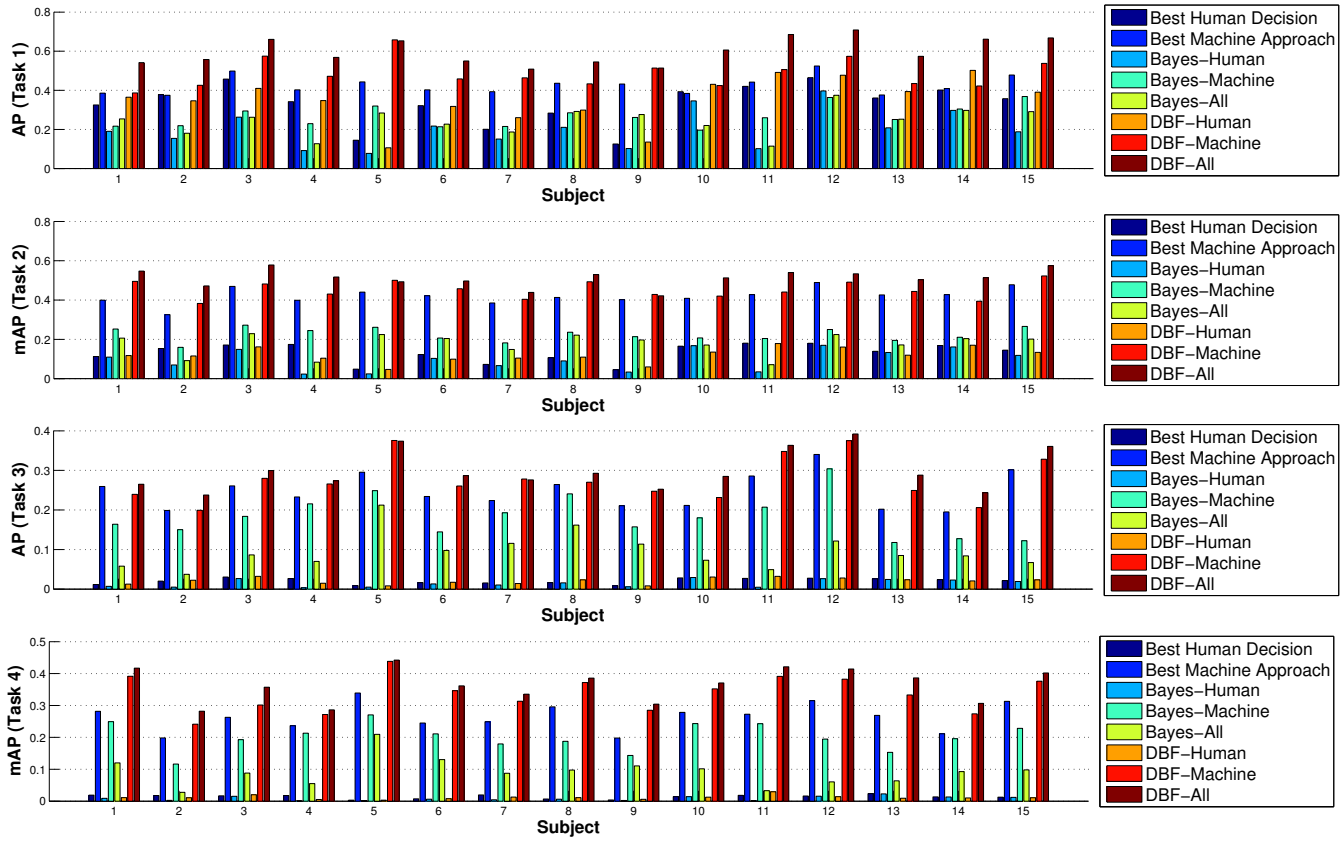

Fig. 10. **Performance of best human and machine approaches as well as fusion approaches per-subject**

for DBF. This supports the key concept behind DBF, that modeling an intermediate state to represent "uncertainty" in detector outputs can be beneficial. The fact that these results are consistent across all tasks suggests that the structure of tasks themselves do not affect the outcome.

The results of each task cannot be directly compared to one another; for instance, Task 2 is more complex than Task 1, thus, it is understandable that AP/mAP will be lower. However, it is interesting to see that DBF-Machine and DBF-All in Task 4 (target-specific) outperform Task 3 (target-agnostic). This may be because the task-conversion algorithm from Task 4 to 3 did not adequately convey the relative certainty between target types (e.g., a score of 0.27 for a poster target may not actually represent the same confidence level as a score of 0.27 for a chair target). The 0.02∼0.03 mAP performance loss during task-conversion of certain individual approaches (ESVM and DPM) caused a significant performance loss in fusion (0.07 for DBF-Machine and 0.08 for DBF-All). Pairwise t-tests indicated a statistically significant difference ($p < 0.001$) for each pair of tasks, including Task 3 and Task 4.

To evaluate the effectiveness of fusion between subjects, we compare the 6 fusion approaches against the best human-based classifier and machine-based detector for each of the 15 subjects for all tasks in Figure 10. For all except the fifth subject in all tasks, DBF-All demonstrated the greatest performance. For task 1, DBF-Human outperforms the best human-based classifier 12 of the 15 subjects and DBF-Machine outperforms the best machine approach for all subjects. As suggested by the previous analysis, the combination of human and machine information using Bayesian fusion underperforms the best individual human and machine approaches for all subjects. The comparisons in Task 2,3, and 4 shows similar tendency as in Figure 10.

One possible confound in using mAP as the metric for comparison is that the measure of precision does not incorporate the number of "misses" for a given classifier. For an object detection task (tasks 3 and 4), this type of metric is appropriate; however for a classification task (Tasks 1 and 2) the misses are often just as important as the accuracy of hits. Thus, to verify that the performance differences we saw using mAP for the classification tasks were not simply a product of the chosen metric, Tasks 1 and 2 were also evaluated using Area Under the ROC Curve (AUC), a conventional metric for classification tasks. A comparison of AUC for all individual and fused approaches is shown in Figure 11. DBF-All outperforms all individual classifier/detector approaches as well as all fusion approaches, as was observed using the AP/mAP metric. DBF-Machine and DBF-All fusion also outperform all versions of Bayesian fusion. A one-way ANOVA test was performed on these results with fusion approach as the main effect, obtaining $[F_{(13,182)} = 35.54,\ p < 0.001]$ and $[F_{(13,182)} = 66.65,\ p < 0.001]$ for Tasks 1 and 2, respectively. Subsequent multiple-comparisons tests demonstrated that

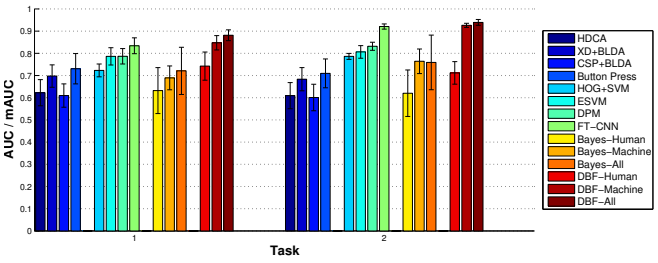

Fig. 11. **Performance comparison with AUC:** Error bars denote the standard deviation across subjects.

DBF-All yields statistically superior mean performance over all approaches except DBF-Machine and FT-CNN (the best computer-vision approach). Note that DBF-All outperforms all other approaches in 13 of 15 and 11 of 15 subjects for Tasks 1 and 2, respectively, indicating a strong trend towards the superiority of DBF-All. We conclude that generating DBF likelihood models provides superior fusion performance based on a combination of the mAP and AUC performance metrics.

V. Conclusion

We have shown that combining Rapid Serial Visual Presentation (RSVP) with computer vision algorithms in a fusion framework can substantially improve target recognition performance, using the precise localization and identification capabilities of computer vision in combination with the precise detection capabilities of a human operator. We have specifically leveraged perception feedback detected from an EEG signal in combination with button-press response, applying positive detections to four state of the art computer vision approaches. Additionally, we have implemented a unique task conversion protocol which overcomes some of the limits of human detection via RSVP, namely the lack of localization and identification feedback.

The treatment of incomplete results for Task 2 as being equivalent to Task 1 has the effect of artificially deflating performance numbers for the human subject response. However, fusion via DBF is still able to leverage unique information from both sources in order to increase overall detection performance. These results suggest that human performance in high cognitive workload situations may be enhanced by integration with fusion approaches, including a novel dynamic belief fusion (DBF) approach. Future work will further explore the effect of randomization of image sequences across subjects and the effects of variable human subject performance over time.

VI. Acknowledgement

This project was supported by the U.S. Army Research Laboratory under a Director's Strategic Research Initiative entitled "Heterogeneous Systems for Information Variable Environments (HIVE)" from FY14-FY16. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] R. B. Girshick, "Fast-RCNN," in *ICCV*, 2015.
[2] J. Touryan, G. Apker, B. J. Lance, S. E. Kerick, A. J. Ries, and K. McDowell, "Estimating endogenous changes in task performance from EEG," *Frontiers in neuroscience*, vol. 8, 2014.
[3] L. Xu, A. Kryzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. on Sys., Man, and Cyber.*, vol. 22, no. 3, may/june 1992.
[4] H. Lee, H. Kwon, R. M. Robinson, W. D. Nothwang, and A. M. Marathe, "Dynamic belief fusion for object detection," in *WACV*, 2016.
[5] M. C. Potter, and E. I. Levy, "Recognition memory for a rapid sequence of pictures," *Jour. of Exper. Psych.*, vol. 81, no. 1, pp. 10–15, July 1969.
[6] M. C. Robeck, and R. R. Wallace, "The psychology of reading: an interdisciplinary approach," *Lawrence Erlbaum Associates*, Hiilsdale, New Jersey, 2nd edition, 1990.
[7] C. B. Mills, and I. J. Weldon, "Reading text from computer screens," *ACM Comp. Surv.*, vol. 19, no. 4, ACM Press, 1987.
[8] S. Mathan, P. Ververs, M. Dorneich, S. Whitlow, J. Carciofini, D. Erdogmus, M. Pavel, C. Huang, T. Lan, and A. Adami, "Neurotechnology for image analysis: Searching for needles in haystacks efficiently," *Aug. Cog.: Past, Present, and Future*, 2006.
[9] M. Santosh, P. Ververs, M. Dorneich, S. Whitlow, J. Carciofini, D. Erdogmus, M. Pavel, C. Huang, T. Lan, and A. Adami, "Neurotechnology for image analysis: Searching for needles in haystacks efficiently," *Aug. cog.: past, present, and future*, 2006.
[10] "`http : //www.dod.mil/pubs/foi/Reading_Room/DARPA/08 – F – 0799_Neurotechnology_for_Intelligence_Analysts_NIA_2008.pdf`".
[11] L. Fei-Fei, A. Lyer, and C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?," *Jour.l of Vis.*, vol. 7, no. 1, pp. 10, January 2007.
[12] K. K. Evans and A. Treisman, "Perception of objects in natural scenes: is it really attention free?," *Jour. of Exper. Psych.: Human Perception and Performance*, vol. 31, no. 6, pp. 1476–1492, 2005.
[13] S. Branson, C. Wash, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie, "Visual recognition with humans in the loop," in *ECCV*, 2012.
[14] A. Kapoor, K. Graumann, R. Urtasun, and T. Darrell, "Active learning with gaussian processes for object categorization," in *ICCV*, 2007.
[15] A. Holub, P. Perona, and M. Burl, "Entropy-based active learning for object recognition," in *OLC*, 2008.
[16] R. M. Robinson, and H. Lee, M. J. McCourt, A. R. Marathe, H. Kwon, C. Ton, and W. D. Nothwang, "Human-autonomy sensor fusion for rapid object detection," *IROS*, 2015.
[17] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.
[18] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-SVMs for object detetion and beyond," in *ICCV*, 2011.
[19] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *CVPR*, 2014.
[20] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *The Annals of Mathematical Staticstics*, vol. 38, no. 2, 1967.
[21] G. Shafer, "A mathematical theory of evidence." Princeton University Press, 1976.
[22] A. D. Gerson, L. C. Parra, and P. Sajda, "Cortically coupled computer vision for rapid image search," *IEEE Trans. on Nerual System and Rehabilitation Engineering*, vol. 14, no. 2, pp. 174–9, June 2006.
[23] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "xDAWN algorithm to enhance evoked potentials: Appliation to brain computer interface," *IEEE Trans. on Biomedical Engineering*, vol. 56, no. 8, pp. 2035–2043, August 2009.
[24] H. Cocotti, B. Rivet, M. Congedo, C. Jutten, O. Bertrand, E. Maby, and J. Mattout, "A robust sensor-selection method for P300 brain-computer interfaces," *Jour. of Neural Engineering*, vol. 8, no. 1, February 2011.
[25] A. Marathe, V. Lawhern, D. Wu, D. Slayback, and B. Lance, "Improved neural signal classification in a rapid serial visual presentation task using active learning," *IEEE Trans. on Neural Syst. Rehabil. Eng.*, Novenber 2015.
[26] P. Sajda, A. Gerson, and L. Parra, "High-throughput image search via single-trial event detection in a rapid serial visual presentation task," in *First International IEEE EMBS Conference on Neural Engineering*, 2003.
[27] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.
[28] P. F. Felzenswalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. on PAMI*, vol. 32, no. 9, pp. 1627–1645, September 2010.
[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.
[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *arXiv preprint arXiv:1409.0575*, 2014.