This is an author version preprint. The final version is available as:
Jeremy Davis, James B. Haynie, Derek T. Anderson, Cindy L. Bethel, J. Edward Swan II, John E. Ball, Amy Bednar, "Spatial Relationship-Driven Computer Vision Image Data Set Annotation", *IEEE International Joint Conference on Neural Networks* (IJCNN), Padova, Italy, July 18–23, 2022. DOI: 10.1109/IJCNN55064.2022.9892975.

Spatial Relationship-Driven Computer Vision Image Data Set Annotation

Jeremy Davis and James B. Haynie *Dept. of Research and Development*
Babel Street Inc.
Washington, DC, United States
jdavis@babelstreet.com Derek T. Anderson *Dept. of Electrical Engineering and Computer Science*
University of Missouri-Columbia
Columbia, MO, United States
andersondt@missouri.edu Cindy L. Bethel, J. Edward Swan II and John E. Ball *Dept. of Computer Science and Engineering*
Mississippi State University
Mississippi State, MS, United States
cbethel@cse.msstate.edu Amy Bednar *Information Technology Laboratory*
United States Army Corps of Engineers
Vicksburg, MS, United States
amy.e.bednar@erdc.dren.mil

***Abstract*—Modern machine learning (ML) is based to a great extent on supervised deep learning models that require large amounts of labeled training data. While image data sets with annotations exist, the annotations are produced manually and possess relatively simple descriptions. To date, none of the freely available labeled image data sets incorporate spatial reasoning, one of Gardner's nine human intelligences. This article presents a new process with open source tools provided to label imagery based on spatial interactions between image objects and automated reasoning under uncertainty. The resulting annotated data can be used to train new ML/AI algorithms and/or help us better understand existing methodologies.**

***Index Terms*—object detection, signal-to-text, scene understanding, spatial relationships in images**

I. Introduction

Current state-of-the-art computer vision is largely based on deep learning. In part, this is because deep learning models have consistently demonstrated superior performance compared to prior hand-crafted approaches. However, training current generation supervised learning-based networks is a time and data intensive process. In many applications, the amount of data required is where bottlenecks occur. Existing image data sets outside the realm of training object detection or image captioning models are scarce, and such existing annotated data sets were manually curated, which itself is a time consuming and laborious process. Additionally, to date, none of the freely available annotated data sets incorporate spatial information between objects in an image, a valuable piece of information that can be used to infer actions being performed in the scene. As such, the effectiveness of scene labeling systems trained on these data sets will be limited due to the lack of spatial reasoning, which was one of Gardner's nine human intelligences [4]. These data sets are also overly simple, where the majority contain object localization and labels and a few of these data sets contain manually crafted scene description labels, which are also very simplistic labels. Current state-of-the-art signal-to-text (S2T) systems [22] [23] [26] primarily use natural language processing (NLP) techniques to construct a final image annotation. These systems also do not incorporate any type of spatial reasoning and can be prone to generating overly simplistic or erroneous labels.

This research presents the implementation of a S2T system for image annotation, which takes a fuzzy inference system (FIS) approach to image annotations where the inference is driven by spatial reasoning between object tuples in an image. Incorporation of spatial reasoning allows the system to provide more refined information in regards to interactions occurring in the scene. The system is constructed in a modular manner such that any number of object localization and spatial reasoning algorithms may be used as input to the FIS. This research also shows that the S2T system can be used to automate computer vision data set image annotations, which may prove to be an invaluable tool for use when training deep learning networks. The code used to construct the image annotation S2T system is freely available for the reader at https://github.com/jedavis82/scene_labeling.

Fig. 1 details the methodology of the implemented S2T system. The system takes raw images as input and then segments objects via object localization. For this system, the You Only Look Once version 3 (YOLOv3) [15] algorithm was used, but it is important to note that any object localization algorithm, including mask segmentation, can be used. YOLOv3 was chosen because the results were easily shareable via code repositories. Segmented objects are then used as input to compute spatial relationships between object tuples. For spatial reasoning, a combination of the Generalized Intersection Over Union (GIOU) [17] and Histogram of Forces (HOF) [14] algorithms are used. These spatial reasoning algorithms were

chosen because they are both well established, but again, it is also important to note that any number of spatial reasoning algorithms can be used as input to the FIS image annotation system. These spatial relationships, along with object segmentations, are the inputs to the image annotation FIS, which results in the final image annotation for the scene. Sec. II presents previous work related to this paper. Sec. III describes the data used for experimentation. Sec. IV describes the processes that generate the object localization and spatial relationship information for each image in the data set. Sec. V details a fuzzy system that generates image annotations using the object localization and spatial reasoning information as input. Sec. VI provides a discussion of the results obtained during experimentation. Lastly, Sec. VII presents conclusions drawn from this research as well as potential future work.

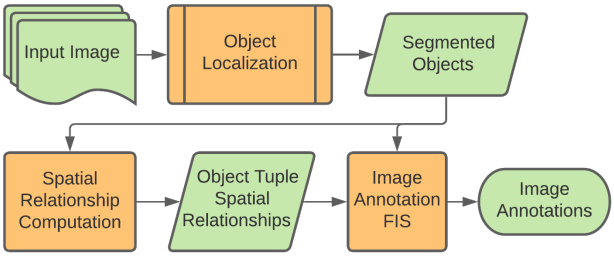


Fig. 1. Step-by-step S2T annotation process pipeline.

II. Related Work

The proposed S2T system constructed in this research makes use of several preexisting algorithms to generate the object localization and spatial reasoning information. YOLOv3 [15] in particular was used for object localization in this work and it is important to note once again that the YOLO model was chosen for easily sharing results in CSV/JSON format. Any number of object localization algorithms can be used with this S2T system, such as the Single Shot Multibox Detector (SSD) [11] bounding box localization algorithm or the Mask R-CNN [5] mask segmentation algorithm.

GIOU [17] was utilized in this work to compute the proximity and overlap relationships between object tuples in an image. Cardinal direction spatial reasoning was obtained by using the HOF algorithm developed by Matsakis et al. in [14], [13], and [12]. Both the GIOU and HOF algorithms were chosen because they were previously well established mathematically and integrate easily into a modular framework based on fuzzy inference as their outputs are essentially fuzzy variables, as discussed in Section IV.

Early works toward image annotation, such as the methods in [9], relied on a hierarchical approach to generate image annotations. While these methods were reliable for annotating data sets on which the methods were developed, the annotations were based on a very strict hierarchy. Thus, such systems were not easily applicable to further data sets outside of the realm of the training data. The work in [1] performed well in generating scene annotations with regards to whether or not a person had fallen and was lying on the ground for an extended state. The S2T system developed in this research follows a similar approach to the methods used in [1] in that the image annotations are generated based around a series of fuzzy rules. While the developed system follows a similar fuzzy rule based approach, the system also incorporates spatial relationship information, along with person-object interactions, into the generated image annotations.

Recent works toward S2T image annotation relied almost exclusively on deep neural network architectures. The works in [22], [23], [7], and [26] for example, all used a convolutional neural network (CNN) to localize objects. Based on the object localization results, a long short-term memory recurrent neural network (LSTM RNN) architecture was then used to construct a semantic description of the most likely actions occurring in the scene. While these frameworks all generated accurate annotations in their individual works, none incorporated spatial relationship information. The S2T system developed in this research aims to provide more informative scene descriptions by incorporating spatial relationship information.

III. Data Set Description

The Microsoft Common Objects in Context (COCO) [10] data set contains depictions of everyday objects in their common contextual settings. 500 images from the COCO 2014 data set were used for the initial development and validation of the S2T system. The COCO 2017 data set was used to perform a sensitivity analysis on the developed system. COCO 2017 is similar to the 2014 data set in that it shows depictions of common objects in their contextual settings. The 2017 data set was chosen for sensitivity analysis because it contained images that the S2T system did not encounter during initial development.

The 2017 data set provides an API that allows a user to filter images based on the types of objects in an image using the FiftyOne [3] library. Images were first filtered to those that contain at least two object segmentation results, and then filtered to those that contain at least one person detection using the FiftyOne library. In total, 2000 images were downloaded from the COCO 2017 data set using the FiftyOne library. These images were filtered such that an image contained at least two object localization results, one of which was a person because the constructed S2T system generates annotations pertaining to people interacting with objects. The methodologies and results presented in the remainder of this paper are based on the 2000 images extracted from the COCO 2017 data set.

IV. Constructing the S2T Inputs

This section describes the processes used to construct inputs for the image annotation S2T system described in Sec. V. These methods include: object localization techniques, metadata computation, proximity and spatial reasoning.

A. Object Localization

YOLOv3 [15] was used as-is, out of the box, to extract bounding box localization results for each of the 2000 input

images. The only caveat is the YOLOv3 non-maxima suppression (NMS) [6] parameter value was chosen as 0.4 via analyzing a random sample of 100 image localization results to tune the parameter such that YOLOv3 retained all valid bounding boxes while eliminating duplicate boxes. While the FiftyOne library also provides bounding boxes and labels for objects in an image, the research performed in this paper relied on computed object localization results in order to simulate an end-to-end S2T pipeline for image data sets where no pre-computed localization results were available. Localization results were filtered such that only images that contained a person and at least one other object were retained. Of the 2000 images, 1637 images contained at least two object localization results, one of which was a person and of the 1637 images, 1243 contained at least one other type of object. Fig. 2 shows examples of localization output from COCO 2017 images that were used as input to the annotation FIS. The localization results were effectively the segmented objects of the image annotation pipeline in Fig. 1.

B. Metadata Computation

Some classes in the COCO [10] data set were vague and non-descriptive, such as the "sports ball" class, which could imply multiple different actions. To overcome such limitations, the Inception [21] model was used to compute metadata for each image in the form of labels, where the top five most likely labels were retained and used as FIS input alongside object localization and spatial reasoning information. Considering the "sports ball" example, Inception labels can be used to switch between the various different sport rules in the FIS.

C. Spatial Reasoning Computation

Spatial reasoning information was the last input required for the FIS and it consisted of two pieces of information for each object tuple in an image: proximity/overlap information and the cardinal direction between the tuple.

Proximity and overlap computation The GIOU [17] algorithm was used to compute both the proximity and overlap relationships. While the original work in [17] measured the accuracy between a predicted and ground truth localization result, the authors denote that the algorithm can also be used to measure the "closeness" of two localized objects, a property exploited by the research of this work. The GIOU score is in the range $-1.0 \leq GIOU \leq 1.0$, where a lesser amount implies two objects are farther away from each other and a higher value implies two objects are in closer proximity to each other.

The GIOU algorithm additionally provides an intersection over union (IOU) score in the range $0 \leq IOU \leq 1$ that indicates the degree of overlap between two objects. GIOU was chosen for this work because its output in effect is a fuzzy variable, which models the uncertainty of proximity and overlap, and it also is mathematically well formulated and validated in the original work of [17]. For the purposes of this research, proximity was modeled as triangular membership functions for the following ranges: very close, close, medium, far, and very far. The overlap relationship was also modeled as triangular membership functions for overlap and no overlap.

The GIOU and IOU outputs and membership functions were modeled such that a human could easily interpret them, providing a meaningful label to an object tuple. The fuzzified GIOU and IOU scores were used as input to the image annotation FIS, and all membership functions are available for free use in the code repository located at https://github.com/jedavis82/scene_labeling. Tab. I provides the proximity and overlap labels applied to the object tuples in Fig. 2, where centroid defuzzification [20] was used to obtain a crisp output from the triangular membership functions for proximity and overlap.

Spatial relationship directionality computation HOF [14] was used to compute the cardinal direction between each object tuple. HOF was chosen because it is a well established algorithm, with strong mathematical proofs in [14], [12], [13]. Thus, HOF is mathematically sound and the proofs provided in the original work indicate the algorithm effectively describes the relative positions between object pairs. HOF also works with both bounding box and mask segmentations, making it very robust and capable of computing output no matter the localization technique used. HOF also provides linguistic capabilities of describing object spatial relationships not only as fuzzy variables usable by a FIS, but also human readable summaries that can be used to provide a meaningful image label for an object tuple.

The HOF algorithm generates three types of force histograms: constant force (F0), gravitational force (F2), and hybrid force histograms. These force histogram outputs were in the range $0 \leq HOF \leq 360$, where the histogram values can be viewed as degrees on the unit circle. The final HOF output was chosen via consensus selection when force histograms agreed, or defaulted to the hybrid output otherwise as the hybrid output contained a mixture of both F0 and F2 outputs.

The force histogram outputs generated by the HOF [14] algorithm were segmented into eight cardinal directions: above, above right, right, below right, below, below left, left, and above left, each represented with a triangular membership function. Fig. 3 shows the resultant force histogram outputs for Figs. 2a - 2f respectively with the caveat that the force histogram shown in Fig. 3f corresponds to the person - tennis racket interaction only. Each figure shows the angle, represented in degrees on the x-axis, and the normalized magnitude of the HOF output on the y-axis from the perspective of how the person is interacting with the object. The triangular membership functions for the cardinal directions are superimposed onto the graph such that the defuzzified output angle can be inferred from the graph.

Prior work in [14], [12], and [13] extensively detailed the modeling of output as fuzzy sets. The triangular membership functions in Fig. 3 rely heavily on this prior work, and are presented alongside the actual HOF output for simplicity. For a human-centric description of the linguistic HOF terms generated, the reader is directed to the original work in [13].

Fig. 3a shows that the HOF output for Fig. 2a can be interpreted as the person is below and to the right of the umbrella because the normalized magnitude angle exists in both the

below and right membership functions. Conversely, Fig. 3d shows that the HOF output for Fig. 2d can be interpreted as the person is to the left of the cellphone because the maximum magnitude angle exists in only the left membership function. Tab. I shows the full set of spatial relationship information computed using the GIOU and HOF algorithms for the images in Fig. 2. While the inputs to image annotation were the fuzzy values, centroid defuzzification [20] was used to obtain crisp outputs for proximity, overlap, and direction in the table.

Object localization and spatial relationship computations were performed for each of the 2000 images. These computations generated, for each object tuple pair in an image: the object bounding boxes, their corresponding class labels, the proximity distance value between the object tuple, an indicator of objects' overlap (both computed using the GIOU [17] algorithm), and a force histogram output detailing in which cardinal direction the objects were with respect to each other (computed using the HOF [14] algorithm). These results were stored in both CSV and JSON files for easy retrieval and usage, and these files are available via the Github repository at https://github.com/jedavis82/scene_labeling. Tab. I shows that the output of the HOF and GIOU algorithms both provide meaningful information for an object tuple. The combination of the GIOU and HOF algorithms are what implement the "Spatial Relationship Computation" stage of the pipeline in Fig. 1.

V. Image Annotations

Image annotations were computed based on the concept of domains. In this research, domains are meant to describe the context in which the image annotations are based. This research focuses specifically on the person domain, in which image annotations are meant to describe interactions between people and objects in a scene.

The person domain FIS implementation was based around a conjunctive system of rules. This research makes use of "human-in-the-loop training" to construct the rule base for the person domain FIS where the authors acted as domain experts. The resulting FIS was modeled as a Mamdani [18] fuzzy control system using the SKFuzzy [25] library for rule aggregation. Fuzzy outputs were defuzzified using centroid defuzzification [20] to obtain the crisp output image annotation for each object tuple. Alg. V shows the process used to construct image annotations.

Algorithm 1 Image Annotation Computation

```
Set T = hierarchically ordered image object tuples
for t ∈ T do
    Set p_t to the proximity score (GIOU)
    Set o_t to the overlap score (IoU)
    Set sr_t to the spatial relationships (HOF)
    Set p_t, o_t, and sr_t as the FIS input
    Set y_t to aggregate of the conjunctive rules
    Set y_t^* to the centroid defuzzified value of y_t
    Assign y_t^* as the annotation of t
end for
```

Alg. 1 shows that the FIS operated on each object tuple in an image. Proximity, overlap, and spatial relationship information for each tuple were set as the input to the FIS. The fuzzy output variable was then computed as the aggregate of the conjunctive rules, at which point centroid defuzzification was applied to obtain the crisp output image annotation label. This process was performed for every object tuple in an image, and for every image in the data set. The remainder of this section focuses on the FIS used to compute image annotations beginning with a brief discussion of the hierarchical ordering of objects.

A. Hierarchical Ordering of Objects

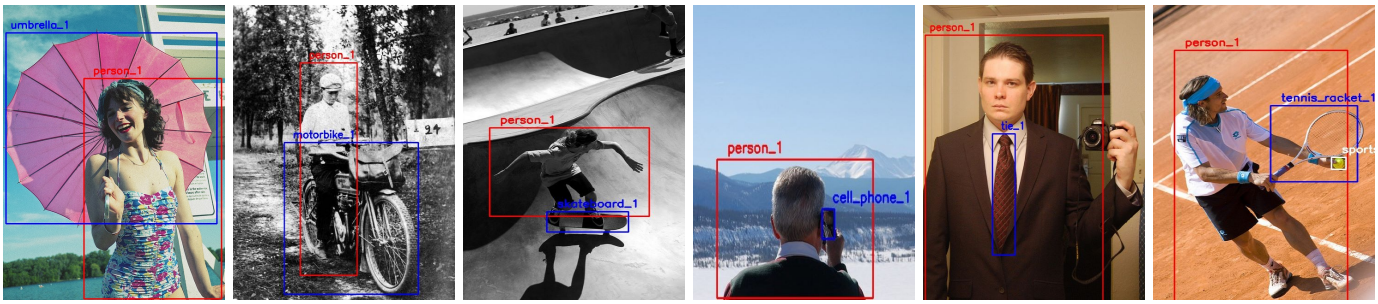
To construct the rules, the authors elected to segment each of the 96 possible COCO [10] 2017 object classes into a hierarchy of 10 categories to group like objects, where each category could contain sub-categories depending on possible actions. These 10 top level categories were: animals, appliances, clothing, electronics, food, furniture, household items, sports, urban, and vehicles. Additionally, object tuples in an image were ordered in a manner where the "person" took precedent over all other objects, followed by animate objects, followed by inanimate objects. This was a design decision made by the authors in order to segment the objects in a manner that allowed for the annotations output by the FIS to follow a consistent pattern.

As an example of an object hierarchy, consider the vehicle category which was broken down into two subcategories: personal and passenger. The personal category contained the COCO objects: car, truck, bicycle, motorcycle, motorbike, and boat. The passenger category contained the COCO objects: airplane, bus, and train. These categories segment objects into subcategories where a human is likely performing the same action with each object in the category. For example, the action for all objects in the personal category was likely "riding/driving" and the action for all objects in the passenger category was likely "riding on/in". For a full listing of the object hierarchy, the reader is directed to the source code at https://github.com/jedavis82/scene_labeling.

B. Person Domain Annotations

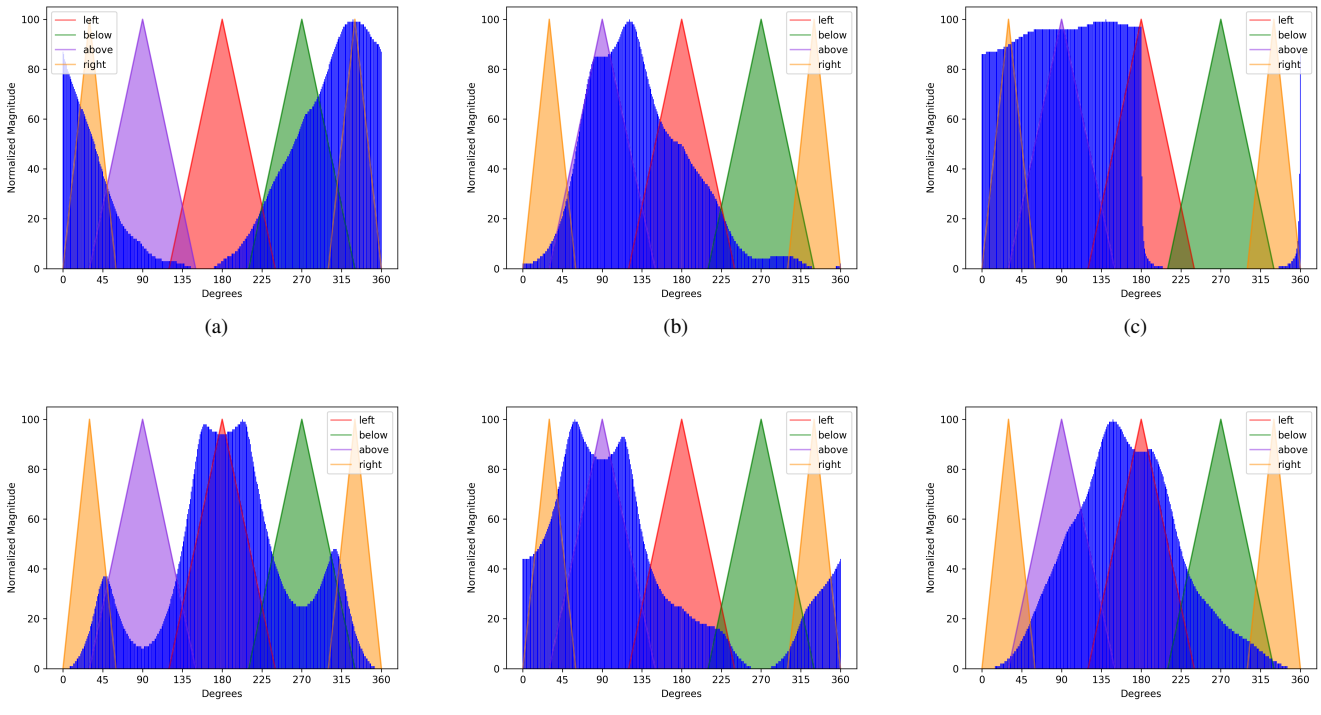
Person domain annotations center around a person interacting with another type of object and serve to provide insight toward what specific actions a person is performing with an object. The person domain required an extensive rule base, and for brevity, the reader is directed to the source code at https://github.com/jedavis82/scene_labeling for a full listing of the person domain FIS rule base used in this research.

To construct the FIS, the first requirement was enumerating all actions that a person-object tuple could perform. This was accomplished using the object hierarchy categories, where each category was constructed to give some insight into what actions a person would likely perform with the object. The authors, acting as domain experts, analyzed each category to determine the set of possible actions for each category and sub-categories. After enumerating the actions, the rule base



(a) (b) (c) (d) (e) (f)

Fig. 2. Object localization results.



(a) (b) (c)

(d) (e) (f)

Fig. 3. Histogram of Forces output.

TABLE I
SPATIAL RELATIONSHIP INFORMATION

Image	Object Tuple	Proximity	Overlap	Direction
Fig. 2a	person_1 - umbrella_1	Very Close	Yes	Below Right
Fig. 2b	person_1 - motorbike_1	Very Close	Yes	Above Left
Fig. 2c	person_1 - skateboard_1	Close	Yes	Above Left
Fig. 2d	person_1 - cellphone_1	Close	Yes	Left
Fig. 2e	person_1 - tie_1	Very Close	Yes	Above Right
Fig. 2f	person_1 - tennis_racket_1	Close	Yes	Above Left
Fig. 2f	person_1 - sports_ball_1	Close	Yes	Above Left
Fig. 2f	tennis_racket_1 - sports_ball_1	Close	Yes	Above Left

was constructed such that the commonalities between object categories could be leveraged to reduce the rule base size.

As an example of the commonalities between like objects, consider again the vehicle category that consists of personal and passenger vehicles discussed in Section V-A. The personal vehicle subcategory objects all were described by the "riding/driving" rule and the passenger vehicle subcategory objects were described by the "riding in/riding on" rule. As such, the seven objects in the top-level vehicle category were described by only two rules.

As mentioned previously, the SKFuzzy [25] library was used to implement the rule base and FIS used for image annotation. The person domain was constructed as a combination of multiple FIS based around the object categories, where each category was constructed as an individual FIS. During runtime, the object class the person was interacting with was used to reference the corresponding object category and appropriate FIS. This was a design decision based around the SKFuzzy library in that it was more efficient to construct each category's FIS once and use the object category to switch between different FIS instances as opposed to applying a single FIS, which would have a much larger rule base, to a person-object interaction. While the system can generate negative interaction labels, e.g "person not riding motorcycle", the authors elected to omit the negative interaction labels to keep the annotations for each image as succinct as possible. Section VI presents a discussion of example system results. All FIS implementation code and the corresponding object category rule bases are available to use freely in the repository located at https://github.com/jedavis82/scene_labeling.

C. Construction of Additional Domains

This research was focused on providing a proof of concept system that demonstrates the capability of automatically generating image annotations. It is important to note that the proposed system was designed to be as flexible as possible. With this in mind, the system can be constructed with any number of localization or spatial reasoning algorithms. Additionally, the system can be constructed for any domain, as long as interactions can be described by a fuzzy rule base that encapsulates the interactions of interest. In other words, this system is not strictly limited to the "person" domain, rather, the "person" domain was chosen simply because the COCO data set was person-centric and no complex expert knowledge was required to construct rules.

Additionally, the goals of this research were to provide a conduit for additional domain constructions outside of the realm of objects contained in the COCO data set. The code provided at https://github.com/jedavis82/scene_labeling serves as a template for constructing such domains. This research demonstrates that a domain-centric FIS can be constructed using a very minimal number of samples, and Section VI shows that these FIS generalize well to additional data sets.

VI. Results

This section presents example annotations generated by applying the domain FIS to the 1243 images of the COCO [10] 2017 data set. These 1243 images contained at least one person detection and one other object type localization as described in Sec. IV. This section also describes a manual validation used by the authors to determine system efficacy. Tab. II presents image annotations for the images of Fig. 2.

TABLE II
Image Annotations

Figure	Person Domain Annotation
Fig. 2a	person_1 carrying umbrella_1
Fig. 2b	person_1 riding motorbike_1
Fig. 2c	person_1 riding skateboard_1
Fig. 2d	person_1 talking on cell_phone_1
Fig. 2f	person_1 playing tennis with tennis_racket_1
Fig. 2f	person_1 playing tennis with sports_ball_1

Tab. II shows that for Fig. 2a, the system detected that the person was carrying the umbrella. For Fig. 2b, the table shows that the system detected the person was riding the motorbike. Tab. II shows that person_1 in Fig. 2c is detected as riding a skateboard. In Fig. 2d, the system generated the annotation of a person talking on a cell phone. Tab. II shows an interesting property for Fig. 2f. Two annotations were generated, one that shows the person is playing tennis with a tennis racket, but the more interesting annotation results from the "sports ball" object class. The system also generated the annotation that the person was "playing tennis with sports_ball_1" and this highlights the capability of the system to generate valid labels by incorporating Inception [21] meta data when the object localization result is vague.

The rule base, fully available at https://github.com/jedavis82/scene_labeling, inherently incorporates the spatial relationship information into the image annotation process. This S2T system, then, is capable of generating not only the image spatial relationship summaries in Section IV, but also incorporates this information into the resultant image annotations because the rules are constructed based around the spatial relationships of the interacting objects. As an output, the S2T system provides to the reader, in both CSV and JSON formats, the following information for each object tuple in an image: the objects' class labels and segmentations, the proximity and overlap scores, the three force histogram outputs (F0, F2 and Hybrid), a spatial relationship annotation, and a person domain image annotation. Results for the 1243 viable images of this research are freely available to the user in the code repository.

System Validation To assess the system, a manual inspection process was performed on a subset of the S2T system results. As there were no baseline annotations, the authors elected to perform a manual inspection of the annotations to ensure that the results accurately reflect the interactions occurring in the scene. A random sample of 500 images of the 1243 images were chosen for inspection.

The manual inspection process was accomplished by visualizing the object localization results for an image using the OpenCV [2] library and printing the corresponding annotations to the console for inspection. Tab. III shows the total annotations obtained by applying the S2T system to the 500 sampled images where correct and incorrect totals are

presented. Interaction types consist of positive interactions and negative interactions, where positive interactions imply that a person *is* performing some action with an object. Conversely, negative interactions imply that a person *is not* performing an action with an object. Tab. III shows that 2029 annotations were generated, 544 of which were positive interactions and 1485 of which were negative interactions. Of the 544 positive interactions, 496 of these were deemed correct, and of the 1485 negative interactions, 1461 were deemed accurate. In total, 72 out of the 2029 person domain annotations were incorrect during analysis. Tab. IV shows that of the 72 incorrect person domain annotations, seven were due to bad localization results, 34 were caused by a depth problem, and five were caused by an occlusion problem, all of which are discussed in Section VII. Therefore, only 26 of the 2029 person domain annotations could not be explained by known system limitations and were considered a bad S2T system result.

TABLE III
ANNOTATION TOTALS

Interaction Type	Correct	Incorrect
Positive	496	48
Negative	1461	24

TABLE IV
INCORRECT ANNOTATION STATISTICS

Total Incorrect	72
Bad Localization	7
Depth Problem	34
Occlusion Problem	5
Bad S2T Result	26

Inspection of the results in Tabs. III and IV for the image annotations show that the system was able to achieve an accuracy of 97% for the 500 sampled images in the person domain. During the analysis, all of the person domain annotations that could not be explained by known system limitations were observed to be in the "grey area" of the FIS output. That is, both the proper annotation and improper annotation had a degree of membership for the corresponding fuzzy membership functions, but the improper annotations were the eventual output of the FIS after centroid defuzzification [20].

Lastly, it is important to once again denote that the 2000 COCO [10] 2017 input images used for this experiment were images that were not encountered during initial development of the S2T system. Thus, the 500 sampled images used for analysis were also not encountered during the construction of the original S2T system and corresponding rule base. It is also worth noting once again, that none of the localization or spatial reasoning algorithms used are set in stone, thus any combination of algorithms plus any combination of rules can be used to generate image annotations, a design decision made in order to make the system easily extendable to additional domains. Given the results of the manual analysis of 500 randomly sampled images, it is a reasonable deduction that the S2T system does generalize well to additional data sets.

VII. CONCLUSIONS AND FUTURE WORK

While the initial goal of the S2T system was image annotation, this work expected to extend that work to the additional task of automating the data annotation process while still providing a S2T system that can be used as a standalone image annotation tool. The S2T system developed successfully incorporated spatial reasoning into image annotation labels, allowing the generation of more informative scene descriptions. The research presented in this paper shows that the S2T system can be extended to additional data sets, previously unknown to the S2T system, with very minimal effort. Results obtained by applying the S2T system, shown in Section VI, provide evidence to support that this system can be used to generate accurate image annotations in the "person" domain for the COCO [10] 2017 data set.

Known System Limitations During experimentation, the authors observed three known limitations of the S2T system developed in this research. The first known limitation hinges on the ability of the object localization model to correctly identify and localize objects in the image. While the YOLOv3 [15] model is efficient in terms of execution speed, it sometimes generates poor localization results, a problem inherent in all current object localization algorithms. Improving object localization model accuracy is beyond the scope of this research, but future iterations of the S2T system should rely on increasingly accurate localization models being developed in the field. Because of the nature of localization algorithms and their potential to generate erroneous results, the system leveraged the HOF algorithm and fuzzy reasoning to help model the inherent uncertainty. While it does not alleviate all of the issues, as results show, it does help to allow the system to be more flexible in terms of generating annotations compared to methods that do not allow uncertainty modeling.

The second observed system limitation occurred when depth estimation was necessary in order to accurately describe the proximity and overlap values. The constructed S2T system works only in the 2D space, and as such, the models used for localization and spatial relationship computation do not provide any estimates of object depth. Without this z value, it is difficult for the system to determine the appropriate measure of proximity and overlap. With this in mind, the system can generate false positive results indicating that two objects are interacting that a human could inherently understand were not interacting. By incorporating algorithms that perform spatial reasoning in 3D, such as [16], or object localization in 3D, such as methods described in [24], the system could extend relatively easily to the 3D space.

The third observed system limitation occurred when an object was occluded by another object in the scene. For example, consider the case where a person's legs are occluded by a horse and all that can be seen is the person's torso which is above the horse. The system, having no knowledge that the person's bounding box is occluded by the horse's bounding box, would likely infer that the person was riding the horse. Human intuition would tell us that the person is not riding the horse, but was simply occluded by the horse, but there is currently no way in the 2D realm to relay this information to

the S2T system. This limitation again, would likely be solved by incorporation of depth information in most cases and as such, future work will focus extensively on the incorporation of depth information to the S2T system.

One final caveat of the person domain S2T system is that it does not apply labels for person-person interactions. These interactions were omitted due to lack of information. The object localization and spatial reasoning algorithms used in this research did not incorporate pose estimation into their output. Pose estimation is a vital component of determining a person-person interaction and as such, the person-person interaction annotations are reserved for future research.

Future Work As future work, research will be performed to alleviate the known system limitations discussed. This research will focus primarily on the incorporation of depth information to the system to improve accuracy. As the FIS does not require 2D or 3D data, but simply object localization and spatial reasoning information as input, applying the system in the 3D realm compared to the 2D realm should prove to be an easily achievable research exercise. The work by Kaur et al. in [8] presents one potential avenue of incorporation of depth information into the S2T system. Additionally, future work will focus on a method of incorporating pose estimation into the S2T system such that person-person interactions can be generated by the S2T system. Lastly, the proposed system only works on pairs of objects. Future research will investigate methods such as [19] to allow the system to extend computation to n-tuples of objects in an image as opposed to only tuple pairs.

REFERENCES

[1] Derek Anderson, Robert H. Luke, James M. Keller, Marjorie Skubic, Marilyn Rantz, and Myra Aud. Linguistic summarization of video for fall detection using voxel person and fuzzy logic. *Computer vision and image understanding : CVIU*, 113 1:80–89, 2009.

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[3] Jason Corso and Brian Moore. Voxel51 // developer tools for ml.

[4] Howard Gardner. *Frames of Mind: the Theory of Multiple Intelligences*. Basic Books, New York City, NY, USA, 15th edition, 1983.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[6] Jan Hendrik Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. *CoRR*, abs/1705.02950, 2017.

[7] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, April 2017.

[8] Jaspinder Kaur, Tyler Laforet, and Pascal Matsakis. Fast fourier transform based force histogram computation for 3d raster data. In Maria De Marsico, Gabriella Sanniti di Baja, and Ana L. N. Fred, editors, *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2020, Valletta, Malta, February 22-24, 2020*, pages 69–74. SCITEPRESS, 2020.

[9] Li-Jia Li, Richard Socher, and Fei-Fei Li. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, pages 2036–2043. IEEE Computer Society, 2009.

[10] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, and Scott E. Reed. Ssd: Single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[12] Pascal Matsakis, James M. Keller, Ozy Sjahputera, and Jonathon Marjamaa. The use of force histograms for affine-invariant relative position description. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(1):1–18, 2004.

[13] Pascal Matsakis, James M. Keller, Laurent Wendling, Jonathon Marjamaa, and Ozy Sjahputera. Linguistic description of relative positions in images. *IEEE Trans. Syst. Man Cybern. Part B*, 31(4):573–588, 2001.

[14] Pascal Matsakis and Laurent Wendling. A new way to represent the relative position between areal objects. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(7):634–643, 1999.

[15] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

[16] Jameson Reed, Mohammad Naeem, and Pascal Matsakis. A first algorithm to calculate force histograms in the case of 3d vector objects. In Maria De Marsico, Antoine Tabbone, and Ana L. N. Fred, editors, *ICPRAM 2014 - Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods, ESEO, Angers, Loire Valley, France, 6-8 March, 2014*, pages 104–112. SciTePress, 2014.

[17] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression, 2019.

[18] Timothy J. Ross. *Fuzzy Logic with Engineering Applications*. Wiley, New York City, NY, USA, 3rd edition, 2010.

[19] Grant Scott, Matt Klaric, and Chi-Ren Shyu. Modeling multi-object spatial relationships for satellite image database indexing and retrieval. In Wee-Kheng Leow, Michael S. Lew, Tat-Seng Chua, Wei-Ying Ma, Lekha Chaisorn, and Erwin M. Bakker, editors, *Image and Video Retrieval*, pages 247–256, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[20] Michio Sugeno. An introductory survey of fuzzy control. *Information Sciences*, 36(1):59–83, 1985.

[21] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[23] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *CoRR*, abs/1609.06647, 2016.

[24] Yilin Wang and Jiayi Ye. An overview of 3d object detection. *CoRR*, abs/2010.15614, 2020.

[25] Josh Warner, Jason Sexauer, scikit fuzzy, twmeggs, Alexandre M. S., Aishwarya Unnikrishnan, Guilherme Castelão, Felipe Arruda Pontes, Tobias Uelwer, pd2f, laurazh, Fernando Batista, alexbuy, William Song, The Gitter Badger, Roberto Abdelkader Martínez Pérez, James F. Power, Himanshu Mishra, Guillem Orellana Trullols, Axel Hörteborn, and 99991. scikit-fuzzy/scikit-fuzzy: Scikit-Fuzzy version 0.4.1, March 2019.

[26] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.