Assigning Relative Importance to Scene Elements

Igor L. O Bastos, William Robson Schwartz
Smart Surveillance Interest Group, Department of Computer Science
Universidade Federal de Minas Gerais, Minas Gerais, Brazil
igorcrexito@gmail.com, william@dcc.ufmg.br

***Abstract*—The human brain is able to rapidly understand scenes through the recognition of their composing elements and comprehension of the role that each of them plays. This process, related to human perception, impacts in what people care when they see an image and the priority they give to each element. The idea of priority, also referred as importance, is based on biological features of perception and social aspects that interfere in how people perceive what they see and what is considered relevant. In this context, this paper proposes the *Element Importance Relative Assignment* (EIRA), an approach that models how humans attribute importance to elements in a scene. This approach is based on perceptual, compositional and contextual features employed to assign importance to elements in a scene. To evaluate the proposed approach, tests are conducted in different image datasets with emphasis on the UIUC Pascal Sentence Dataset, where our approach achieves an average accuracy of 86.89%.**

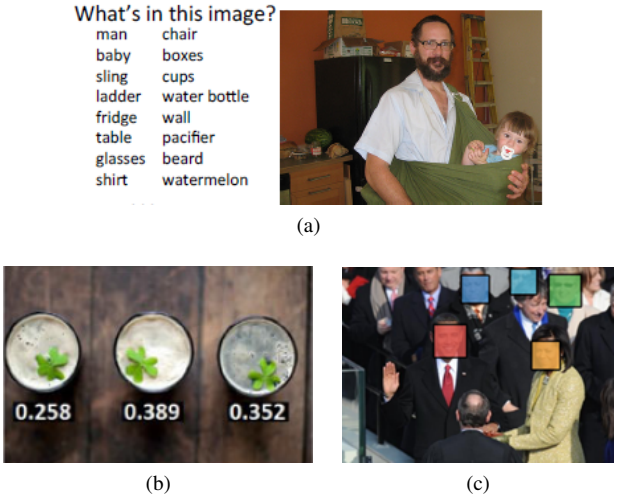
I. Introduction

The scene understanding performed by humans is a semantic process in which meaningful and informative descriptions are extracted from images [1]. In this process, the concept of *importance*, highly related to context, plays a relevant role [2]–[5] and has been subject of the Computer Vision community, being usually modeled as a combination of social rules and biological features [6]–[9].

The study of importance has been investigated by different approaches, ranging from techniques exploiting local image information [10] to those with focus on the relationship between elements that compose a scene [8], [9], [11], e.g., people, animals or any object that is relevant for the scene understanding. The approach proposed in this work, named *Element Importance Relative Assignment* (EIRA), belongs to the latter category.

Similarly to EIRA, Berg et al. [7], Kong et al. [8] and Mathialagan et al. [9] proposed approaches to compute an importance score to elements in images. Through their research, Berg et al. [7] studied the impact of compositional features (e.g. scale, position, distance to other elements) and semantic features in the task of importance assignment. They presented an approach to discover the probability of an object to be mentioned in a scene. On the other hand, Kong et al. [8] discarded the idea of category, presenting a method to assign an importance score to similar elements in a same image, using object characteristics, such as area, contrast, saliency and focus as their main features.

Focusing specifically on assigning importance to people in images, Mathialagan et al. [9] proposed an approach called VIP, which aims at predicting importance of individuals in group photographs. VIP considers a number of objects features, encompassing compositional features such as distance and scale, and aspects to deal specifically with images of people, such as face pose estimation and face occlusion [9]. Figure 1 presents images and outputs obtained with the approaches of Berg et al. [7], Kong et al. [8] and Mathialagan et al. [9].



(a) (b) (c)

Fig. 1. Approaches tackling element importance assignment. (a) Importance based on mentioning order [7]; (b) score assignment for similar objects [8]; (c) people importance scoring [9].

Also related to the present approach, Hwang and Grauman [6] investigated the concept of importance associated to the image retrieval task, by exploiting the relationship among how humans tag images, the relative importance of objects and their layout in a scene. Aiming at boosting queries, image objects are used as features that describe the context, besides compositional and semantic aspects, such as their position and the order in which they are tagged by humans.

In addition to the aforementioned work, others modeled importance of elements, such as Spain and Perona [1] and Yu et al. [12]. Although still working with the idea of importance, those researches focus on the assembling of importance models. Spain and Perona [1] developed a model to associate human observers' labels to keywords that describe a scene and Yu et al. [12] developed a model related to attention, coding local objects through features as saliency and colors, associated to the behavioral Gestalt theory.

Focused on aspects of human perception and social rules,

TABLE I

FEATURES EMPLOYED BY DIFFERENT IMPORTANCE APPROACHES. NUMBER OF ELEMENTS (NE), OCCURRENCE VECTOR (OV), CATEGORY (CAT), ENVIRONMENT (ENV), SCALE (SC), DISTANCE METRICS (DIST), SALIENCY (SALI), FOCUS (FOC), SHARPNESS (SHARP), DEPTH (DEP), DEEP EXTRACTED APPEARANCE (DA), GAZE (GAZ), FACE OCCLUSION (FO), POSE ESTIMATION (PE). *ENVIRONMENT DESCRIPTION AND OCCURRENCE VECTOR ARE ANNOTATED BY USERS IN BERG ET AL. APPROACH [7]

	Contextual				Compositional		Perceptual				Appearance	People specific		
Methods	NE	OV	CAT	ENV	SC	DIST	SALI	FOC	SHARP	DEP	DA	GAZ	FO	PE
Kong et al [8]					✓	✓	✓	✓						
Berg et al [7]		✓*	✓	✓*	✓	✓								
Mathialagan et al [9]					✓	✓			✓			✓	✓	✓
proposed EIRA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			

(a) (b)

Fig. 2. Importance of elements in the images. (a) person is more important than horse; (b) sofas are equally important.

this paper proposes the method called *Element Importance Relative Assignment* (EIRA), responsible for assigning relative importance to elements in images, generating an importance order. Although similar to approaches found in the literature, mainly those proposed by Berg et al. [7], Kong et al. [8], and Mathialagan et al. [9], our method employs a richer set of features and presents the advantage of being more general in the sense that it can be applied to datasets with multiple categories, differently from the work proposed in [8] and [9]. The proposed approach combines multiple features, creating a vector composed of focus, saliency, sharpness and depth responses, object characteristics (dimension, position, quadrant, category), environment description, and appearance features. Table I lists the features employed by several approaches. It is important to notice that our approach contemplates a larger set of features than the other approaches, allowing us to capture more information regarding the relative importance of the elements in the scene.

According to experimental results, EIRA achieves an average accuracy of 86.89% on the $1,000$ test images of the UIUC Pascal Sentence database [13], a dataset composed of 20 classes. To validate this accuracy, statistical tests were conducted, evidencing the good results obtained with EIRA and the contribution of its features for the final accuracy.

II. PROPOSED APPROACH

The proposed approach, the *Element Importance Relative Assignment*, consists in a method to assign relative importance to elements that compose an image (e.g., objects and people) through pairwise scores, i.e., each pair of elements receives a score indicating either the most important element or an equal importance for both. After computing this importance for every pair of elements in an image, an importance ranking is obtained, as shown on Figure 2, where the indexes on the bottom left corner of the bounding boxes show the order (lower rankings indicate higher relative importances).

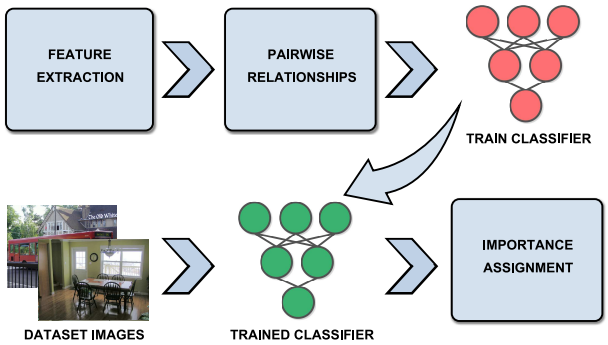


Fig. 3. Steps of the proposed approach.

Since the focus of our proposed approach is the relative importance, it is adequate to be applied in images containing multiple objects. For each pair of objects in the image, features are extracted and an importance relationship is composed. The importance relationships represent an exhaustive combination of object pairs and correspond to the feature vectors used for training/testing the approach. To do that, it is also necessary to associate an importance label to each vector. Figure 3 illustrates the methodology applied in EIRA and next sections detail the approach.

A. Feature Extraction

Researches focused on importance assignment are commonly based on two main factors: human perception and social rules [7]. While the former is driven by the visual attractive elements in images and can be modeled considering aspects such as saliency and focus, the latter is weighted by values of human society and by the way people usually consider some scene elements more important than others based, for instance, on their category (e.g., people tend to be more important than dogs and cars tend to be more important than roads).

To model human perception and social rules, which enables the generation of feature vectors that approximate the human importance behavior, we consider aspects such as *perceptual*, *compositional*, *contextual* and *appearance features*, with the last being computed using deep learning approaches.

1) Perceptual Features: The perceptual features correspond to a set of characteristics associated to how humans notice elements in the scene, approximating a biological process

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

Fig. 4. Maps generated from the perceptual techniques employed in EIRA. a) original image; b) saliency map; c) focus map; d) sharpness map; e) depth map.

[14]. In the present approach, this process is modeled through the generation of saliency, focus, sharpness and depth maps. Saliency, focus and sharpness features are employed to generate an score, which weights the response inside element boxes compared to the rest of image. This score is normalized by the area of the object. In turn, the depth response is obtained through the computation of the median (50th-percentile) of the object distance, representing the central point of the depth distribution. The choice for the median regards its good central tendency representation even if the data is skewed, what is important in the cases that the element pixels present a high distance variation. Figure 4 illustrates outcomes of the saliency, focus, sharpness and depth techniques for an image of the UIUC Pascal Sentence database.

Saliency. Since the importance is also related to the way people perceive an object in an image, it is necessary to cover this point on the process of importance assignment. Looking at a scene without prior knowledge, drives our attention to some locations, mostly because of their saliency, defined by contrasts in color, intensity, orientation [15]. This perception employs a mechanism to narrow the incoming data, focusing on parts with relevant information, even in environments with high clutter and multiple objects [16]. The saliency algorithm employed in EIRA is the one proposed by Luo et al. [17], which handles the saliency as a function of the color variance and its distribution on the image.

Focus. The focus feature regards the goal of the photographer when the picture was taken. Different focus values in the image indicate the existence of zones (focused) that tend to be more important than others (blurred). To model that, we employed the algorithm based on blur application and response of a gradient kernel on different regions of the image proposed by Zhuo and Sim [18].

Sharpness. The sharpness feature follows the same principle of the focus. The addition of this feature intends to enforce the distinction between blurred and focused zones. To gather a sharpness score, the gradient response on a region of the image is computed. The algorithm employed in EIRA is the same employed by Mathialagan et al. [9].

Depth. Estimating depth is an important component to understand a scene. The closer an element is to the camera (photographer), the more important it tends to be. To model this aspect, the deep convolutional network model proposed by Eigen et al. [19] is used. This network receives a RGB image as input, outputting the corresponding depth map. It is worth mentioning that the network weights and its architecture were provided by Eigen et al., requiring, therefore, no additional training.

2) Compositional Features: Compositional features are related to how elements are arranged in an image [6], [7]. This arrangement is considered by human viewers in the process of image understanding and importance assignment, with emphasis on two main factors: scale and placement (position) of elements, both used in the present approach.

Scale. This feature is associated to the idea that objects closer to the camera tend to be more important than the ones farther and therefore, they appear in a larger scale. Following this simple statement, scale is computed as the area of the element normalized by the area of the image.

Placement. Placement features regard the position of an element in the image. These features are associated to the idea that photographers tend to center what they want to show and therefore, elements in a central position are usually considered more important than the ones placed on corners. In EIRA, the placement of an element derives from the distance between its centroid and the image center. Besides that, four different distance metrics are employed: (i) distance to image center normalized by the largest dimension of the element, (ii) normalized distance to the average centroid (considering the centroid of all elements), (iii) normalized distance to weighted centroid, in which the area of each element is taken into account to determine the average centroid, and (iv) element quadrant.

3) Contextual Features: As supported by Spain and Perona [1], the context plays an important role for the task of image understanding and importance assignment. In this sense, contextual features are employed in EIRA with the aim of describing the scene in which elements are inserted.

Number of elements. This feature computes the number of bounding boxes annotated in each image, which is associated to an idea of context.

Category. The category feature brings a description of the element. The labels 'person' or 'dog', for example, attributed to each person or dog boxes, correspond to the category of these elements. Spain and Perona [1] and Berg et al. [7] pointed the relevance of the category for the importance

Fig. 5. Responses from the environment description network. For each scene, the top-1 prediction is exposed along its probability.

assignment task, as humans tend to attribute more importance to some categories when compared to others.

Occurrence vector. The occurrence vector is represented by an array composed by 20 positions (one position for each category). Each array slot stores the number of elements of the corresponding category. For instance, if an image contains two people, the slot array relative to person receives the value 2, and the rest of the slots receive 0. This counting tries to tackle the context in which every element is inserted.

Environment description. To describe the environment represented by the image, the convolutional network developed and trained by Zhou et al. [20] is considered. This network outputs 365 probability responses associated to different places. With that, it is possible to characterize the environment represented by the scene, providing an information that could be useful for the assignment of importance. One can think that the environment affects the importance of an element. For instance, in a zoo picture, animals are much likely to be photographed and probably deserve a higher importance score. Figure 5 presents some outputs obtained with this network for images of the VIP Dataset [9], one of the datasets in which the present approach was tested.

4) Appearance Features: The purpose of using these features to model importance is related to the fact that people tend to consider objects to be more important when they have a certain appearance. For instance, a person in frontal pose tends to be considered more important in an image than a person in profile, as illustrated in Figure 6.

The feature vector used in our approach is also composed by object appearance features, which are extracted with the employment of the deep Alexnet network [21] pre-trained in the ImageNet database. In EIRA, each object is presented to this network and appearance features are extracted from the activations of the convolutional layer *fc7*, for which, according to experiments, the approach presented the best results.

B. Pairwise Relationships

For each image containing more than one element, feature vectors are composed comprising two objects at once. These vectors contain information regarding the image, the two elements and a label to indicate which element is more important or if they are equally important. It is worth mentioning that for each image, all possible combinations of elements are

Fig. 6. Appearance associated to the importance. A person in a frontal pose tends to be more important than in profile.

covered in this process. Figure 7 depicts the structure of the feature vector assembled for each combination of elements in an image. Since contextual features are mostly related to the images, they are used to characterize them, excepting for the category. The category feature typifies each element and so, is placed among the element features.

C. Importance Assignment

The last step of EIRA consists of the importance assignment. The data are split in three classes, related to the possible labels that can be associated to each pairwise relationship. For the data classification, tests were conducted considering different classifiers and a multilayer perceptron was selected since it presented the best results. Thus, for each pair of objects in an image, this classifier outputs their relative importance, evidencing whether they are equally important or one is more important than the other. After computing this importance for each pair of elements (exhaustively combined in the image), the importance order related to the whole image is obtained.

III. Experimental Results

To evaluate the present approach, experiments were conducted considering the UIUC Pascal Sentence [13] and the VIP dataset [9]. For both, cross-validation protocols were employed to gather the accuracy of EIRA, allowing us to compare the approach with other methods proposed by Berg et al. [7] and Mathialagan et al. [9]. The experiments are detailed in the next sections.

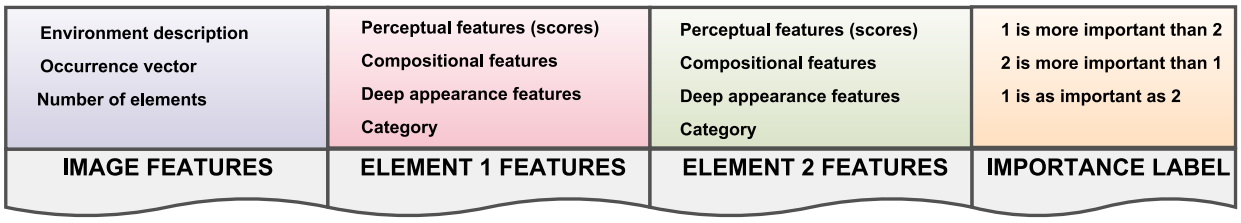


Fig. 7. Feature vector assembled from the combination of each pair of elements in the image.

A. Datasets

The proposed approach is intended to assign an importance order to elements in a scene. In this sense, the UIUC Pascal Sentence [13] and VIP [9] datasets were used with the aim of evaluating our method.

UIUC Dataset is composed by $1,000$ images and $3,430$ elements associated to 20 different categories. Since our approach computes the relative importance of elements, images containing a single object are discarded. Besides that, it is important to mention that for training and testing EIRA, pairwise relationships were composed. This way, considering the discarded images, $10,575$ feature vectors were assembled, representing an exhaustive combination of elements in every image. For the UIUC Pascal Sentence, a 5-fold cross validation was employed, following the protocol executed by Berg et al. [7], state-of-the-art for this dataset.

The VIP dataset is composed by 200 group of people images. For this dataset, 3720 pairwise relationships were composed and a 10-fold cross validation protocol was employed, as described by Mathialagan et al. [9], state-of-the-art method for the dataset.

B. Importance Annotation

As the UIUC Pascal Sentence is a dataset related to object recognition applications, it does not contain importance labels, making it necessary to generate values to represent the output of each pairwise combination. In the case of VIP dataset, it was necessary to annotate each person face of every image and so, ask users to annotate importance labels, since this dataset also does not provide such information.

We developed a tool that allows users to perform importance annotations for pair of elements. This tool provides an interface that assigns importance labels to every pair of annotated objects (annotation of objects is provided by the dataset), in which users are presented each pair of elements and asked to establish an importance order between them.

The idea behind the annotation tool is very similar to the one presented by Mathialagan [9]. For the present research, 4 different users were asked to perform the annotations over all images of the dataset and a majority voting was employed to determine each relation label[1]. Ties are treated with a random choice between the most voted labels. Figure 8 shows the interface of the annotation tool employed in the present research. The right buttons A, B and $=$ are used to determine the relation labels.

[1]The importance annotations created are available in the following link: http://www.ssig.dcc.ufmg.br/importance-assignment/.

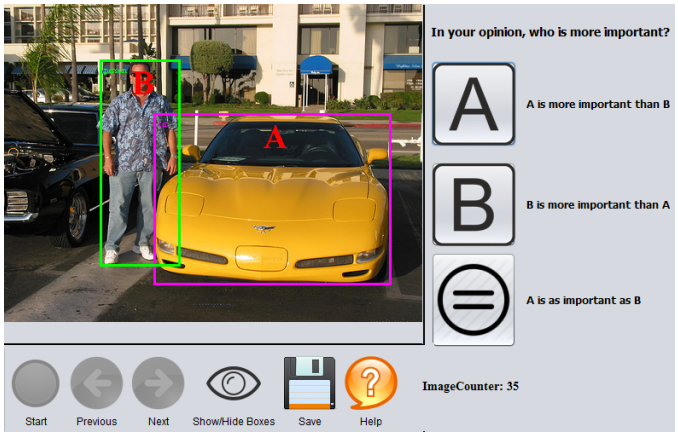


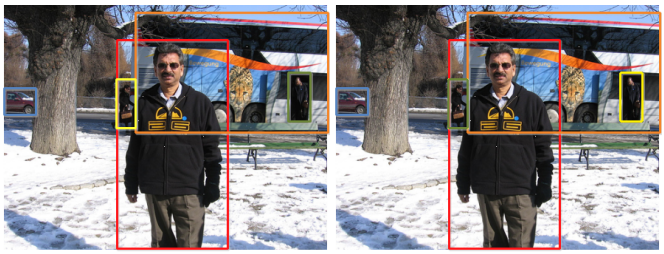
Fig. 8. Interface of the annotation tool.

C. Evaluation on the UIUC Pascal Sentence Dataset

To evaluate EIRA, a first experiment was conducted on the UIUC Pascal Sentence dataset. With that, an average accuracy of 86.89% was obtained, highlighting a good performance of the approach and the discriminability of the feature vector employed.

For a better understanding of the results, the divergence among users' annotations was also computed, which allows the acquisition of a weighting coefficient for each relation of the dataset. This coefficient considers the level of agreement of users' labels. For instance, if all four users agree, the weight given to this pairwise relationship is maximum (1). However, if three users agree, the weighting coefficient is 3/4 of the maximum (0.75), and so on. With that, an additional test was conducted considering different weights for each relationship. A weighted accuracy of 89.96% was obtained.

Our last experiment on UIUC Pascal sentence dataset considered a protocol similar to the one proposed in [9], where a cross-human agreement was computed. This protocol considers the use of one user annotations as labels and, considering the remaining three, a majority voting is applied. The accuracy is computed comparing the annotations used as labels and the outcomes of the majority voting. This process is repeated until all users have their annotations used as labels and it measures the agreement among users for the importance assignment task. The obtained accuracy was 89.14%, indicating that not



(a) (b)

Fig. 9. EIRA importance output and corresponding ground-truth. The color of the boxes indicate the importance, the reddish colors show higher importances. (a) EIRA importance output; (b) annotated importance ground-truth (better viewed in colors).

TABLE II
ACCURACIES ACHIEVED WITH THE PROPOSED APPROACH.

Approach	Accuracy (%)
EIRA	86.89
EIRA (weighted)	89.96
Cross-Human agreement	89.14

even humans were able to perform this task with a complete agreement and that EIRA's results are not far from it. Table II summarizes the outcomes from the application of EIRA in the UIUC Pascal Sentence dataset.

A visual inspection of EIRA results allows us to see that the approach performed well even for complex scenarios. Figure 9 shows the output obtained with EIRA for one image of the UIUC Pascal Sentence Dataset and its corresponding ground-truth. It is noticeable that EIRA output is similar to the ground-truth, excepting for one box. This erroneous outcome can be associated to the fact the confounded elements (green and yellow boxes) appear in a very similar scale, depth and with similar appearance. However, the yellow boxed object, annotated by users with lower importance, is placed in a more central position, causing the approach to incorrectly attribute a higher importance.

D. Comparisons

We compare our approach to the ones proposed by Berg et al. [7] and Mathialagan et al [9]. Although the non-availability of the annotations made by Berg and differences regarding the objective of the approaches, both were applied to the same set of images - UIUC Pascal Sentence - and are related to the idea of element importance assignment using our created annotations. In addition, to compare to the approach of Mathialagan et al. [9], it was necessary to annotate element boxes (people) and importance labels for the VIP Dataset [9], as aforementioned.

Table III presents the obtained results with EIRA and the approach of Berg et al. [7] for the UIUC Dataset. The method proposed in [7] achieved an accuracy of 82.00% for the UIUC Dataset while EIRA obtained 86.89%.

TABLE III
COMPARISON BETWEEN EIRA AND UIUC PASCAL SENTENCE STATE-OF-THE-ART METHOD.

Approach	Acc. (%)
Berg et al. [7]	82.00
EIRA	**86.89**

TABLE IV
COMPARISON BETWEEN EIRA AND VIP STATE-OF-THE-ART METHOD.

Approach	Acc. (%)
Mathialagan et al. [9]	92.72
EIRA	**88.46**

Table IV shows the accuracies of EIRA and Mathialagan methods for VIP dataset. EIRA obtained an accuracy of 88.46%, which is lower than the 92.72% reported by Mathialagan et al. [9]. However, this accuracy cannot be directly compared to the results achieved in [9] since the object annotations were not provided neither the importance labels. Another point to be highlighted is that the approach proposed in [9] is specific for people importance scoring, containing features related to that purpose and not being applicable in different contexts, such as the UIUC Pascal Sentence database. It is worth to note that the approach proposed by Berg et al. [7] relies on the idea of category and it is not suitable for a 1-category dataset as VIP.

E. Cross-Dataset Results

Due to the accurate results obtained with EIRA for both datasets, an experiment was conducted to verify how well the approach would behave considering one dataset for training and the other one for test. First, the UIUC Pascal Sentence database was used for training the approach and the VIP dataset for testing, achieving an accuracy of 68.01%. Then, considering the VIP for training and UIUC Pascal Sentence for test, the obtained accuracy was 44.69%. The poor results obtained for both datasets can be associated to the feature responses for each.

On the case of the Pascal Sentence for training and VIP for test, the dataset was trained in Pascal Sentence and a high weight was driven to the category feature (this dataset presents 20 different categories). This feature, pointed by Spain and Perona [1], Hwang and Grauman [6] and Berg et al. [7], as one of the most important for importance labeling, was not considered on the test performed on VIP database, since this dataset contains elements from a single category (person).

On the case of the VIP for training and UIUC Pascal Sentence for test, the results were even worse. This fact can be associated to the category label, which was not weighted by the classifier since the training was performed on VIP. In addition, the extracted appearance features are not suitable for the UIUC Pascal Sentence since the training process only extracted appearance responses from people and not other objects.

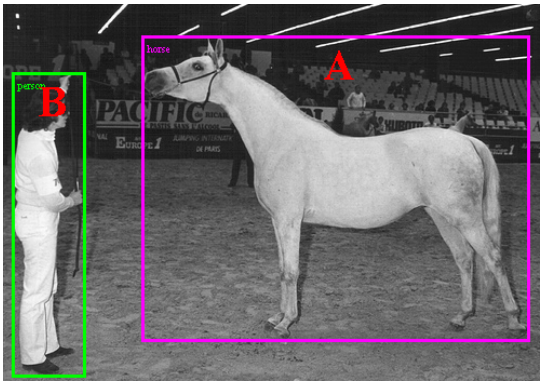


Fig. 10. Relationship annotated with a high degree of divergence.

F. Influence of Annotation Subjectivity

The proposed approach presented accurate responses for both datasets where it was applied. However, it is important to consider points that may threaten the validity of the obtained results, as the subjectivity of the annotation process, which may vary according to the opinion of each person and is intensified by the complexity of the databases and the number of users asked to label the datasets.

The inherent subjectivity of the labeling process leaded to divergences among users annotations, which were noticed for different images of both datasets. The scene presented in Figure 10, for instance, was annotated with a high degree of divergence. For the person and horse denoted by A and B, two users annotated both elements with equal importance, one annotated the horse (A) as the most important and one annotated the person (B) as the most important. This divergence emphasizes the subjectivity of the process and the difficulty to annotate importance relations, even for humans. One point that needs to be mentioned, in this sense, is that this subjectivity could have dropped the accuracy of the approach, since a same user could have annotated similar cases with different labels.

A high number of user annotations could contribute to reduce the impact of the subjectivity on the labeling process. However, annotating every image of the datasets is a very costly task, since for each of these images, all possible combinations of elements need to be labeled.

G. Statistical Analysis and Feature Contribution Evaluation

Despite the high accuracies obtained with EIRA, it is necessary, as a final step of the approach, a statistical study regarding the obtained results. This study was conducted through the repetition of the training and classification steps 15 times in the UIUC Pascal Sentence Dataset. The 86.89% accuracy value reported during the paper corresponds to the average value obtained. Besides that, it was obtained an standard deviation value of 2.02 and a coefficient of variation of 0.0232, indicating a low dispersion of the data around the mean value.

After computing these initial metrics, the confidence interval for the mean value was determined using the *t-distribution*. As the experiments were reproduced 15 times, it was necessary to guarantee the normality of the data, evidenced by the *Anderson-Darling* test [22], [23]. Then, the confidence interval was computed considering a 99% confidence value. As the result, it was obtained the interval $[85.34, 88.44]$. The low interval range obtained for a high confidence value is justified by the low dispersion of the data, evidenced by the coefficient of variation. With that, it is assumed that for 99% of the experiments, an accuracy between the interval margins tends to be obtained.

TABLE V
ACCURACY OBTAINED FOR EACH VARIATION OF THE FACTORIAL DESIGN.

Appearance	Perceptual	Compositional	Avg Acc. (%)
-1	-1	-1	63.90
+1	-1	-1	76.27
-1	+1	-1	69.79
+1	+1	-1	79.62
-1	-1	+1	71.52
+1	-1	+1	79.56
-1	+1	+1	71.59
+1	+1	+1	86.89

A complete factorial design ($2^k r$) [24] was also conducted to understand the contribution of features for the final classification accuracy. To that end, the feature vector used in EIRA was split into the four groups aforementioned in Section II-A: perceptual, compositional, contextual and appearance features. However, as the contextual features are responsible for describing the image (as depicted in Figure 7) and contain only the category information that could be used to compare two elements, this group was used as the baseline of the approach, not being targeted by the factorial study. Thus, the remaining groups of feature were exhaustively added/removed from the feature vector and the accuracy was computed, as shown in Table V, where -1 indicate that the feature group was removed, while $+1$ indicates that the feature was added. The first line of the table corresponds to the baseline, i.e., the accuracy obtained with the contextual features. It is interesting to see that contextual features provided an average accuracy of 63.90%, since this group of features contains only the category to discriminate the importance of elements.

According to Table V, it was possible to compute the contribution of each feature group and the amount of variation associated to them in relation to the baseline. The appearance features, for example, were responsible for 57.69% of the variation, while the perceptual were responsible for 7.99% and the compositional for 11.45%. It is important to mention that almost no variation was associated to the combination of features. Table VI shows the variation associated to each feature and their combination. One can notice that the total variation explained by the features is 80.52%, meaning that the remaining 19.48% are associated to experimental error and/or divergence among feature contributions inside a same group.

For a more precise study, each feature of each group should

TABLE VI
VARIATION EXPLAINED BY EACH FEATURE AND THEIR COMBINATION.

Features	Variation (%)
Appearance	57.69
Perceptual	7.99
Compositional	11.45
Appearance + Perceptual	0.64
Appearance + Compositional	0.0003
Perceptual + Compositional	0.0009
Appearance + Perceptual + Compositional	2.75
TOTAL	**80.52**

be studied. However, the conducted factorial design seems to be satisfactory for the present paper, as it evidences the contribution order and amount of variation associated to each group of features.

IV. CONCLUSIONS AND FUTURE DIRECTIONS

The proposed approach produced satisfactory results, leading us to believe that the features used for importance prediction are discriminative. Some points can be connected to the erroneous results, such as the complexity associated to the UIUC Pascal Sentence and VIP databases, and consequent divergence in user's annotations. Even with the impossibility to perform a direct comparison of results, it could be seen that the results are comparable to other importance techniques. In addition, the statistical tests performed indicate that EIRA presents a very deterministic behavior, evidenced by the low coefficient of variation and low difference between confidence interval margins. Finally, EIRA shows to be valuable since it is a general model, being applicable in any image dataset.

As future directions, we intend to carry out a more precise and deep factorial design, which would provide more information about feature contribution. We also intend to make the importance annotations available along the guidelines used in the present approach, allowing comparisons of other methods to EIRA.

V. ACKNOWLEDGEMENTS

The authors would like to thank the Brazilian National Research Council – CNPq, the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17) and the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project). The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the GeForce Titan X GPU used for this research.

REFERENCES

[1] M. Spain and P. Perona, "Measuring and predicting object importance," *International Journal of Computer Vision*, vol. 91, no. 1, pp. 59–76, 2011.

[2] C. L. Zitnick, T. Chen, and D. Parikh, "Exploring tiny images: The roles of appearance and contextual information for machine and human object recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 34, pp. 1978–1991, 2012.

[3] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Comput. Vis. Image Underst.*, vol. 114, no. 6, pp. 712–722, Jun. 2010.

[4] S. K. Divvala, D. Hoiem, J. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection." in *CVPR*, 2009, pp. 1271–1278.

[5] O. Marques, E. Barenholtz, and V. Charvillat, "Context modeling in computer vision: Techniques, implications, and applications," *Multimedia Tools Appl.*, vol. 51, no. 1, pp. 303–339, Jan. 2011.

[6] S. J. Hwang and K. Grauman, "Accounting for the relative importance of objects in image retrieval," in *Proceedings of the British Machine Vision Conference*, 2010, pp. 58.1–58.12.

[7] A. C. Berg, T. L. Berg, H. D. III, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi, "Understanding and predicting importance in images," in *CVPR*. IEEE Computer Society, 2012, pp. 3562–3569.

[8] Y. Kong, W. Dong, X. Mei, C. Ma, T.-Y. Lee, S. Lyu, F. Huang, and X. Zhang, "Measuring and predicting visual importance of similar objects," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, p. in press, 2016.

[9] C. S. Mathialagan, A. C. Gallagher, and D. Batra, "Vip: Finding important people in images." in *CVPR*, 2015, pp. 4858–4866.

[10] M. Fink and P. Perona, "Mutual boosting for contextual inference," in *Proceedings of the 16th International Conference on Neural Information Processing Systems*, ser. NIPS'03, 2003, pp. 1515–1522.

[11] A. Torralba, K. P. Murphy, and W. T. Freeman, "Contextual models for object detection using boosted random fields," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds., 2005, pp. 1401–1408.

[12] Y. Yu, G. K. I. Mann, and R. G. Gosine, "An object-based visual attention model for robotic applications," *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, vol. 40, no. 5, pp. 1398–1412, Oct. 2010.

[13] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, ser. CSLDAMT '10, 2010, pp. 139–147.

[14] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Visual Perception, Progress in Brain Research*, vol. 155, 2006.

[15] D. Walther, U. Rutishauser, C. Koch, and P. Perona, "On the usefulness of attention for object recognition," in *in: Workshop on Attention and Performance in Computational Vision at ECCV*, 2004, pp. 96–103.

[16] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artificial Intelligence*, vol. 146, no. 1, pp. 77 – 123, 2003.

[17] S. Luo, Z. Liu, L. Li, X. Zou, and O. L. Meur, "Efficient saliency detection using regional color and spatial information," in *European Workshop on Visual Information Processing, EUVIP 2013, Paris, France, June 10-12, 2013*, 2013, pp. 184–189.

[18] S. Zhuo and T. Sim, "Defocus map estimation from a single image," *Pattern Recogn.*, vol. 44, no. 9, pp. 1852–1858, Sep. 2011.

[19] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, ser. NIPS'14, 2014, pp. 2366–2374.

[20] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, "Places: An image database for deep scene understanding," *arXiv preprint arXiv:1610.02055*, 2016.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1097–1105.

[22] P. Closas, J. Arribas, and C. Fernndez-Prades, "Testing for normality of uwb-based distance measurements by the anderson-darling statistic," in *Proceedings of IEEE Future Network and Mobile Summit*, 2010.

[23] S. Engmann and D. Cousineau, "Comparing distributions: the two-sample anderson-darling test as an alternative to the kolmogorov-smirnoff test," *Journal of Applied Quantitative Methods.*, vol. 6, no. 3, pp. 1–17, 2011.

[24] O. Valle-Casas, R. Dalazen, V. Cene, and B. Alexandre, "Complete factorial design experiment for 3d load cell instrumented crank validation," in *Proceedings of the 37th Annual International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2015.