Pattern Recognition 53 (2016) 174–183

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Understanding image concepts using ISTOP model

M.S. Zarchi [a,*], R.T. Tan [c], C. van Gemeren [b], A. Monadjemi [d], R.C. Veltkamp [b]

[a] Faculty of Engineering, Haeri University, Meybod, Iran
[b] Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands
[c] Yale-NUS College, Singapore
[d] Faculty of Computer Engineering and Information Technology, University of Isfahan, Isfahan, Iran

ARTICLE INFO

Article history:
Received 14 November 2014
Received in revised form
16 October 2015
Accepted 14 November 2015
Available online 4 December 2015

Keywords:
Visual term
Context sensitive grammar
Latent SVM
And–Or graph
Concept recognition
Image parsing

ABSTRACT

This paper focuses on recognizing image concepts by introducing the ISTOP model. The model parses the images from scene to object's parts by using a context sensitive grammar. Since there is a gap between the scene and object levels, this grammar proposes the "Visual Term" level to bridge the gap. Visual term is a higher concept level than the object level representing a few co-occurring objects. The grammar used in the model can be embodied in an And–Or graph representation. The hierarchical structure of the graph decomposes an image from the scene level into the visual term, object level and part level by terminal and non-terminal nodes, while the horizontal links in the graph impose the context and constraints between the nodes. In order to learn the grammar constraints and their weights, we propose an algorithm that can perform on weakly annotated datasets. This algorithm searches in the dataset to find visual terms without supervision and then learns the weights of the constraints using a latent SVM. The experimental results on the Pascal VOC dataset show that our model outperforms the state-of-the-art approaches in recognizing image concepts.

1. Introduction

Image understanding is an open problem in computer vision. Conventionally, objects in an image are considered as the main concepts in image understanding or annotation approaches [1,2]. Thus, to understand a scene, the image is parsed at the object level by the objects detection algorithm. Although by performing an object detection algorithm, some important concepts can be detected, this approach has three main drawbacks. First, concepts at higher levels than objects are ignored [3]. These concepts can be made by combining related objects. For example, consider the scenes 'a' in Fig. 1. Although the objects "horse" and "person" can be detected, the "horse riding" concept cannot be recognized at object level. We can say that there is a significant gap between the scene level and the object level in parsing an image. Second, the constraints among objects such as locations, ability to occlude each other, and aspect ratios are not considered. For example, when a horse is detected, we expect to see a person on the back or standing beside the horse but not below the horse. Third, when objects conduct an action together, they might occlude each other, and it becomes difficult to recognize objects, because occlusions cause deformation in the appearance of the objects.

In order to solve these drawbacks, we introduce a novel model to (i) bridge the gap between objects and scenes, (ii) use constraints to improve object detection and (iii) solve the occlusion problem. Furthermore, this model has the ability to be trained on a weakly annotated dataset, where only bounding boxes of objects are available.

In the proposed model, we introduce a new high level concept between the scene level and the object level in order to bridge the gap between them. We call this new concept "Visual Term". A visual term is the composition of the related co-occurrence objects that represent a higher level concept. In Fig. 1, the visual terms are shown in blue boxes. Since, our model parses an image from the scene level into the visual term, object level and part level, we call it "ISTOP". By analogy to natural languages grammar, each level of this model has its corresponding level in English grammar as shown in Table 1.

The ISTOP model uses a context sensitive grammar (CSG) [4] in parsing an image, which imposes context and constraints. In this grammar, the objects and their parts are detected by filter templates that are trained by a part-based approach. In order to handle the occlusion problem, we determine co-occurrence objects by a data mining approach firstly. Then, we train filter templates for occluded objects in addition to the general filters. Hence, an occluded object can be represented more properly than

* Corresponding author.
E-mail addresses: sardari@haeri.ac.ir (M.S. Zarchi), elertt@nus.edu.sg (R.T. Tan), C.J.Vangemeren@uu.nl (C. van Gemeren), Monadjemii@eng.ui.ac.ir (A. Monadjemi), R.C.Veltkamp@uu.nl (R.C. Veltkamp).

http://dx.doi.org/10.1016/j.patcog.2015.11.010
0031-3203/

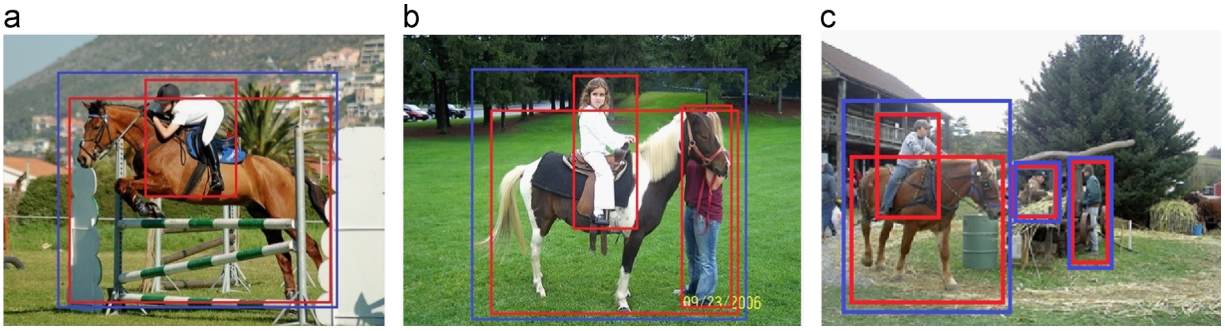

Fig. 1. Three examples of visual terms. The visual terms are shown by blue bounding boxes and their objects by red bounding boxes. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 1
The ISTOP grammar levels and their corresponding levels in English grammar.

English grammar level	ISTOP grammar level
Alphabet	Part: Segment or partition of an object
Word	Object: Representation of a physical object
Phrase/clause (term)	Visual Term: Composition of related objects
Sentence	Scene: Interpretation of the entire image

in the existing models. Fig. 2 shows the occlusion filter templates for a person object occluded by a horse. The overall training phase of the ISTOP model is shown in Fig. 3. The details of the ISTOP model are explained in Section 3.

2. Related works

Many approaches are proposed to detect image concepts, however only a few of them focus on narrowing down the gap between the scene and object level. Here we shortly mention the most relevant ones to our work.

Objects detection is the important tasks in each approach. Recently, discriminative approaches such as bag of visual words [5,6] and discriminative part based model [7,8] have received more attention in object and image classification. The discriminative part based model (DPM) [8] is a well known model for object detection. In this model, each object is described by a variation of the histogram of oriented gradient (HOG). The DPM includes a root filter and a set of part filters as shown in Fig. 2. The filter response is given by convolving each filter to HOG features at a given location. The advantage of this model is that it can consider appearance changes due to view point or gestures by defining multiple filters for an object category. In the extended version of this model [7], the context is used to improve object detection. In this extension, the results of object detection of all object categories are used blindly as context, however this approach has two drawbacks. First, it increases the time complexity and second, to use context for an object, the results of object detection for all the other categories should be prepared.

With respect to parsing images, the grammar based models [9,10,4] are most relevant to our work. In these models, a grammar is used to parse images from high level concept (scene) to low level concept (primitives). For instance, the attribute graph grammar [9] is proposed to represent man-made scenes. This grammar has six production rules, which can generate spatial layout of the detected rectangular surfaces. The main difference of these models to ours is that they need a fully annotated dataset for their training while ours only uses a weakly annotated dataset and can be used for general purposes.

The idea that the relation between objects can be modeled as a higher concept level than the object level has been received more attention recently [3,11,12]. Sadeghi and Farhadi introduce an intermediate concept between the object and the scene as Visual Phrase [3], which is similar to our Visual Term. They consider two related objects as a rigid concept and represent it in one filter by using the DPM, however in our method we create filter templates for each object separately, and the detected objects are allowed to merge together by a predefined grammar. Our proposed visual term has three main advantages over their work. First, our model can be performed on a weakly annotated dataset with all kinds of relations, while in the phrasal model, the authors create a specific dataset for their model and only specific relations between objects is considered. For example, for the horse and person relation, only a person riding a horse is modeled, while we consider all types of relation in our work such as: a person riding, jumping and walking a horse. Second, in our model the position of each object is determined by its filter, while in their work, the object positions are not determined explicitly. Third, in our visual term, multiple objects can be involved. For example, in the relation of a person with a sofa, more than one person can sit on a sofa. In this case, our model detects all persons and sofa and composes them to a visual term.

In some specific domains, the relation between objects and their context is used to improve recognizing objects and interactions [13–16]. Desai and Ramanan [13] present an approach to model human interactions by combining the strengths of articulated skeleton [17], visual phrase [3] and poselet [18] approaches. The mutual context model [15] is proposed to jointly model objects and a human in human–object interactions. In this approach, object detection and human pose estimation are used to improve the accuracy of detecting the objects that interact with the human.

3. The ISTOP model

In our model, a context sensitive grammar is used to parse images. This grammar can be represented in an And–Or graph (AOG). The or-nodes of the graph indicate a selection choice between multiple alternatives by a production rule (E_{or}) as specified in Eq. (1). In this equation, O, A and t are or-node, and-node, and terminal (leaf) node respectively:

$$E_{or} : O_i \rightarrow A_1 | A_2 | \ldots | A_n | t_1 | t_2 | \ldots | t_n \quad (1)$$

The and-nodes of the graph uniquely indicates the child nodes and a set of constraints on them. This production rule (E_{and}) is defined as

$$E_{and} : A_i \rightarrow (\{O_1, O_2, \ldots, O_n\}, C_{A_i}) \quad (2)$$

where C_{A_i} is a set of constraints on children.

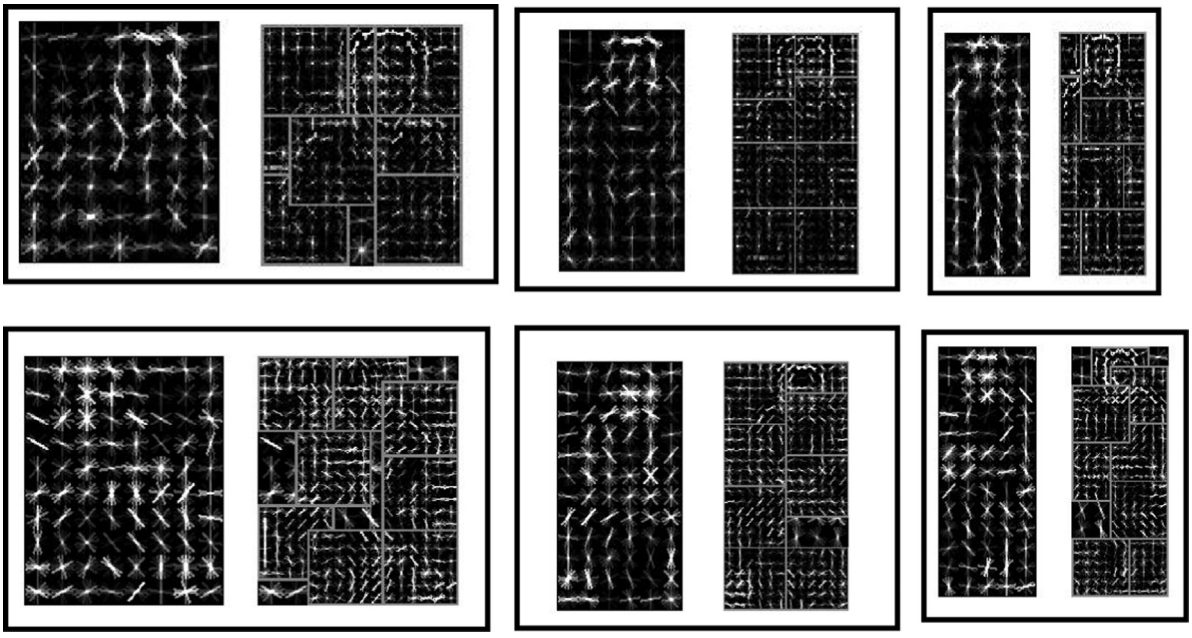

Fig. 2. Trained filter templates for person object and its 8 parts by using discriminative part based model. The first row contains general filter templates for a person. The second row contains the occlusion filter templates for the person when co-occurred with a horse.

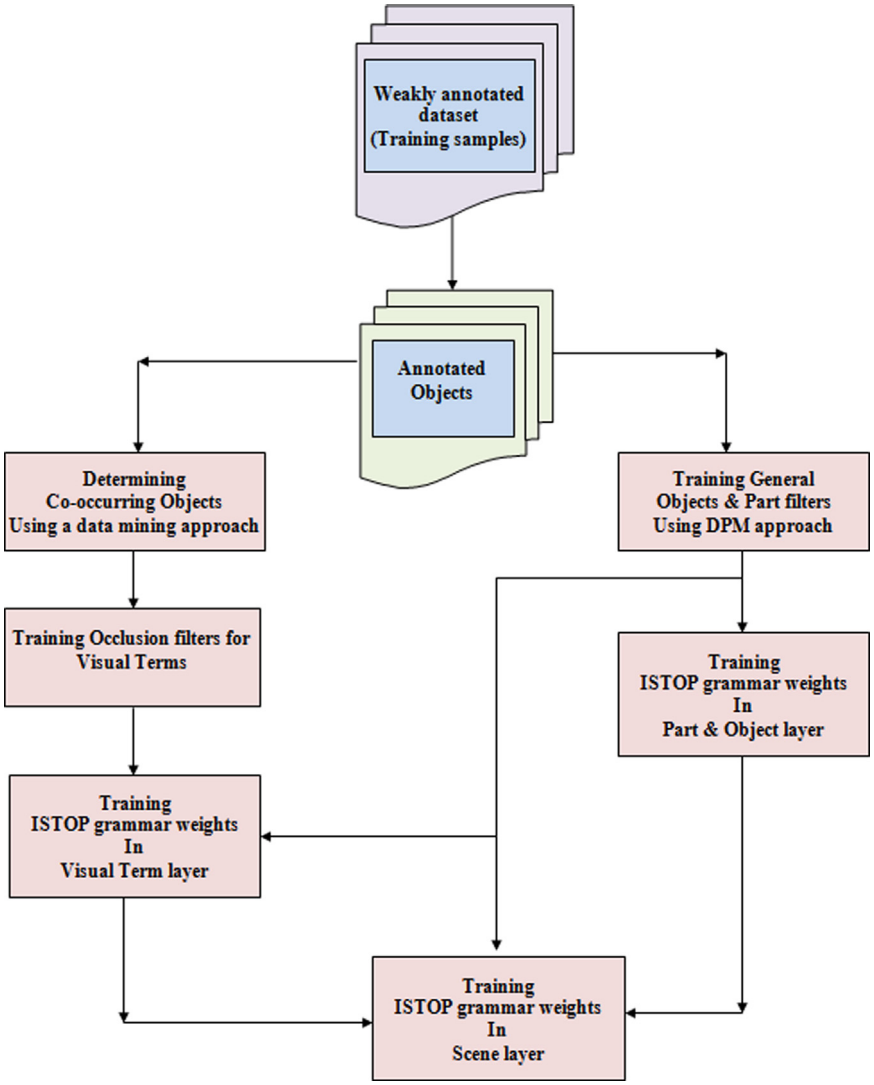

Fig. 3. The overall training phase of the ISTOP model.

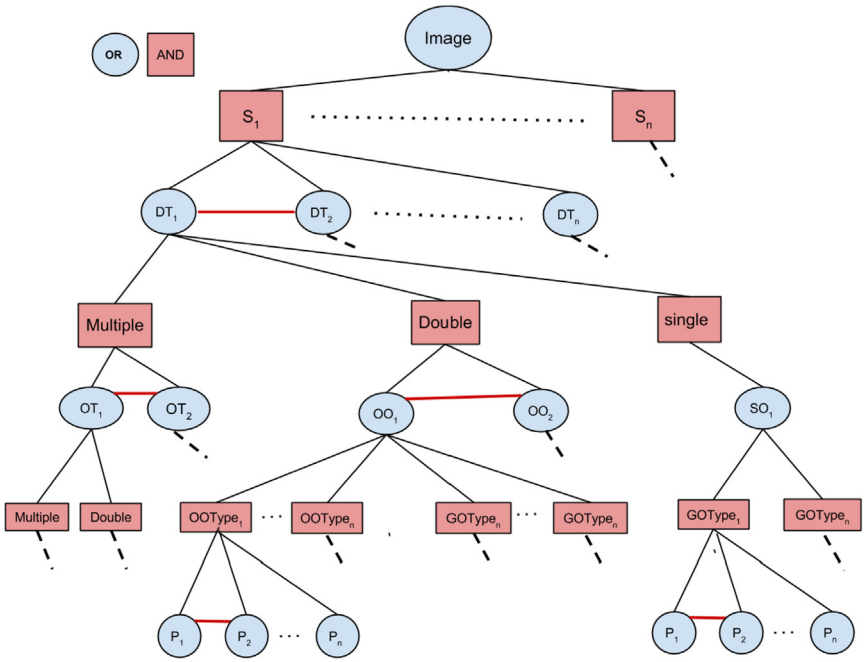

Fig. 4. The And–Or graph of the ISTOP model. The rectangles are and-node and circulars are or-nodes. The horizontal lines are constraints between the children of the and-node and the vertical lines are the composition rules.

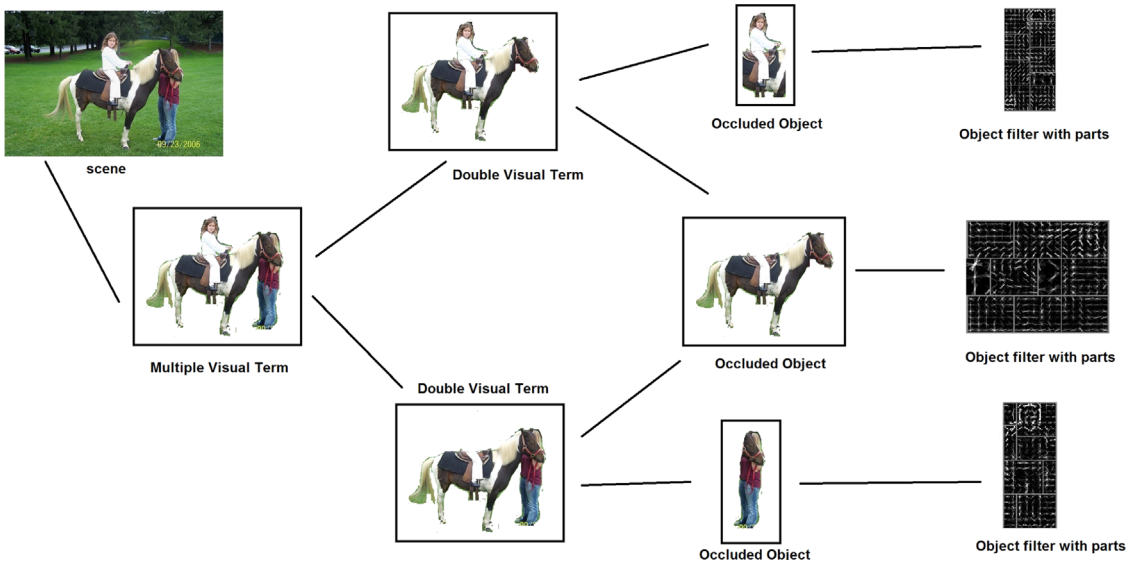

Fig. 5. An example of image parsing by the proposed grammar.

The And–Or graph is defined as in Eq. (3), where $V_{\mathcal{G}}$ is a set of nodes that include non-terminal nodes (and-nodes and or-nodes) and terminal nodes (leaves). E_{and} and E_{or} are the edges to the children of and-nodes and or-nodes. $C_{\mathcal{G}}$ is a set of constraints over the graph. Finally, $O_{\mathcal{S}}$ is the root node of the graph:

$$\mathcal{G} = (V_{\mathcal{G}}, E_{and}, E_{or}, C_{\mathcal{G}}, O_{\mathcal{S}}) \tag{3}$$

The scheme of the proposed And–Or graph is depicted in Fig. 4. In this figure, the horizontal lines represent constraints, and vertical edges show production rules. The production rules and constraints of this graph are explained in Section 3.1. A parse graph is a substitution of graph elements at Or-nodes. In Fig. 5, a pictorial parsed image by this graph is shown. The procedure of parsing an image by using the proposed grammar is explained in Section 3.2. We also refer to Ref. [4] for the fundamental concept of the And–Or graph and image grammar.

3.1. Production rules

The production rules consist of 11 distinct rules that can be explained as follows:

1. $Image \rightarrow S_1 | \ldots | S_n$: Image is the root of the graph and can be parsed into one of the predefined scenes such as “farm”, “street” or “kitchen”.
2. $S_i \rightarrow (\{DT_1 \& \ldots \& DT_n\}, C_{S_i})$: A scene is parsed into $n >= 1$ distinct visual terms (DTs). If $n > 1$, then there are constraints (C_{S_i}) between distinct terms, which are shown by the horizontal

lines in the graph. Here, the important constraint is that DTs do not occlude each other. For example, in Fig. 1, the image 'a' and 'b' have one DT while the image 'c' has three DTs.

3. $DT_i \rightarrow Single|Double|Multiple$: DT can be parsed into "Single", "Double", or "Multiple" terms which have one, two and more than two objects respectively. For instance, in Fig. 1, the DT of image 'a' is "Double", 'b' is "Multiple" and the image 'c' has two "single" and one "double" term.
4. $Single \rightarrow (\{SO\}, C_{Single})$: Single is parsed into a single object (SO). C_{Single} is a constraint on the SO indicating that the SO does not have occlusions with other objects. For example, in image 'c' of Fig. 1, there are two SOs.
5. $Double \rightarrow (\{OO_1 \& OO_2\}, C_{Double})$: Double is parsed into two occluded objects (OOs). Two occluded objects should satisfy a set of ($constraints C_{Double}$) in the grammar such as the percentage of occlusion, aspect ratio and, the relative locations of the two objects. In image 'a' of Fig. 1, the "Double" is parsed into "occluded horse" and "occluded person".
6. $Multiple \rightarrow (\{OT_1 \& OT_2\}, C_{Multiple})$: Multiple is parsed into two occluded terms (OTs) by considering a set of constraints ($C_{Multiple}$). In this rule, the important constraint is that two occluded terms have a shared object. In image 'b' of Fig. 1, the "Multiple" term is parsed into two "Occluded Terms". The first one is "person" on the back of "horse" and the second is "person" standing near "horse" where the shared object is "horse".
7. $OT_i \rightarrow Double|Multiple$: An occluded term can be parsed into "Double" or "Multiple" terms. In this way every "Multiple" term is broken down recursively until it reaches "double" by using rules 6 and 7.
8. $SO_i \rightarrow GOtype_1|\ldots|GOtype_n$: A single object can be parsed into one of the general object types. The general object types are detected by the general filter templates. The general filters for objects are learned by considering all possible training objects. For example, for "horse", three general object filters are learned which correspond to different views of the horse.
9. $GOType_i \rightarrow (\{Gpart_1 \& \ldots \& Gpart_n\}, C_{GOType_i})$: Each general object type is comprised of n parts with regard to the defined constraints (C_{GOType_i}). Since we use the DPM [19] in training, n is set to 8 as shown in Fig. 2.
10. $OO_i \rightarrow OOtype_1|\ldots|OOtype_n|GOtype_1|\ldots|GOtype_n$: An occluded object is parsed into one of the general object types or occluded object type (OOType). The occluded object filters for OOTypes are trained in similar way to general object filters, yet in their training only occluded objects are used instead of all available objects. The reason that we also add general object in this rule as alternative is that sometimes in a "Double" term one object is in background and the other one is foreground. Hence the foreground object does not have deformation and subsequently can be detected better by general filters.
11. $OOType_i \rightarrow (\{Opart_1 \& .. \& Opart_n\}, C_{OOType_i})$: This rule is similar to rule 9, which parses an occluded object types to n parts.

3.2. The parse graph

In order to parse an image, the optimal parse graph for the image is generated by the AOG. A parse graph is an instantiation of the AOG which is defined as

$$pg = (V_{pg}, E_{pg}, C_{pg}) \tag{4}$$

where V_{pg} is the set of node instantiations, E_{pg} show the corresponding composition from E_{or}, E_{and}, and C_{pg} specify corresponding constraints from C_G.

The optimal parse graph is computed by

$$(pg^*|I) = \arg\max_{pg \in \mathcal{G}} Score(pg|I) \tag{5}$$

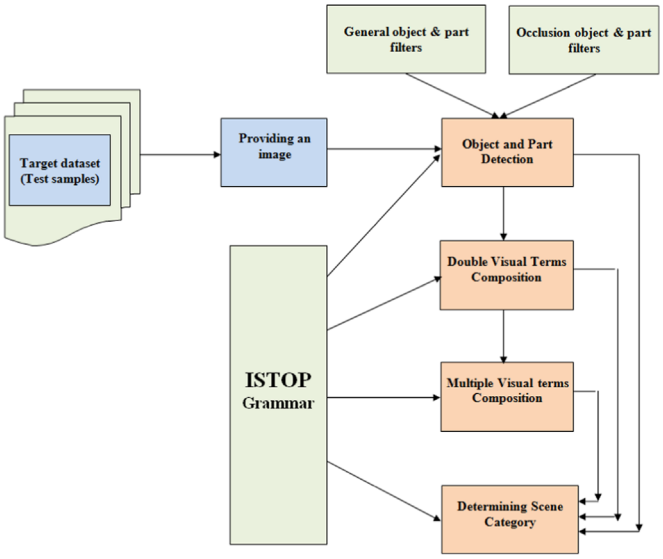

Fig. 6. The block diagram of image parsing by ISTOP model.

where $Score()$ compute constraint satisfaction and $Score(pg|I) = Score(O_S|I)$. This equation indicates that an image can be parsed in different paths (i.e. an And–Or graph can have different parse graphs). Hence, in parsing an image, the path that has the maximum score will be selected by this equation.

Given an input image I, the $Score$ for nodes in G is computed as follows:

1. For each terminal node $t \in V_T$ a filter template trained by using DPM introduced in [7,8]. The filter is convolved at a specified location and the response is considered as the score of the terminal node. This can be defined as

$$Score(t|I) = \theta_t^{app}(\delta), \tag{6}$$

where δ is the position of filter template and θ^{app} computes the filter response.

2. For an or-node $O \in V_{or}$ we have

$$Score(O|I) = \max_{v_i \in ch(O)} Score(v_i|I) \tag{7}$$

which means that the child which has the maximum score is selected on an Or-node.

3. The score of and-node $A \in V_{and}$ is defined as

$$Score(A|I) = \sum_{v_i \in ch(A)} \lambda_i \cdot Score(v_i|I) - \gamma \cdot \Phi(ch(A)) \tag{8}$$

where λ and γ are learned during a training phase by using Latent SVM [20], and $\Phi(ch(A))$ is the constraint penalty for the children of A. This equation indicates that on an And-node, all the children should exists in the path and the summation of their scores and constrain penalty is considered as its score.

In the proposed grammar, we define two kinds of constraints, strict and flexible. The strict constraint determines feasibility of firing a production rule. For example, in rule 4, the strict constraint define that SO does not have occlusion with the other objects or in rule 6, the strict constraint define that two occluded terms should have a shared object. The flexible constraints are introduced to find the best configuration of components in the grammar.

The satisfactions of flexible constraints are measured by a probability density function ($f()$). The probability density function is estimated from the train samples by a histogram based approach ($\hat{f}()$). For example, for rule 5, the percentage of occlusion (how many percent of an object is occluded by the other object),

Fig. 7. The results of performing the ISTOP model on four samples. In each row, the first image is the original image. The second image shows the extracted candidate objects and the third show candidates for double visual terms. The fourth image shows the selected double visual terms by model. Finally, the fifth image shows the merged visual terms to form multiple visual terms.

the aspect ratio of two co-occurrence objects, and the relative locations of the two co-occurrence objects are considered as flexible constraints and the satisfaction of these constraints is measured by an estimated probability density function. Here, the satisfaction constraint implies that the probability of two objects with these specifications (occlusion percentage, aspect ratio and relative location) can form a double visual term. In order to use flexible constraint satisfaction in Eq. (8) we define $\Phi() = 1 - \hat{f}()$ as a penalty constraint.

Table 2
Comparing average precisions of the 12 Visual term categories.

Visual terms	ISTOP	Baseline	Object based
Bicycle with person	**49.9**	45.0	19.9
Boat with person	**3.4**	0.1	0.12
Bottle with dining table	4.2	**4.8**	3.6
Bottle with person	**17.3**	2.9	1.3
Chair with dining table	**18.4**	17.0	8.9
Chair with person	**8.5**	3.8	1.7
Chair with sofa	**6.4**	1.4	2.1
Dining table with person	**19.9**	12.1	10.5
Dog with person	**2.2**	0.5	0.11
Horse with person	**59.6**	56.3	27.6
Motorbike with person	**35.6**	30.4	31.4
Person with sofa	**13.5**	1.3	1.0

Although the proposed And–Or graph is illustrated in a top-down direction, in the implementation of our model, we use a bottom-up approach to create the And–Or graph based on the production rules, then we find the optimum parse graph as we explained in this section. Namely, the objects and their parts are first detected by the general and occluded filters, then all candidate double visual terms are constructed and finally, candidate multiple visual terms are formed. Among all these candidates, those place in the optimum parse graph are selected as the components of the parse image. A brief overview of this procedure is depicted in Figs. 6 and 7.

3.3. Visual Terms in weakly annotated datasets

In a weakly annotated dataset, only presence or absence of an interested object is indicated by a bounding box [21]. Hence, to find visual terms, we use a data mining approach. The visual terms are the related objects that appear together frequently in the images and occlude each other. In order to find the object categories that can form the visual term, a weighted co-occurrence matrix (W) for all object categories is created. Similar to the

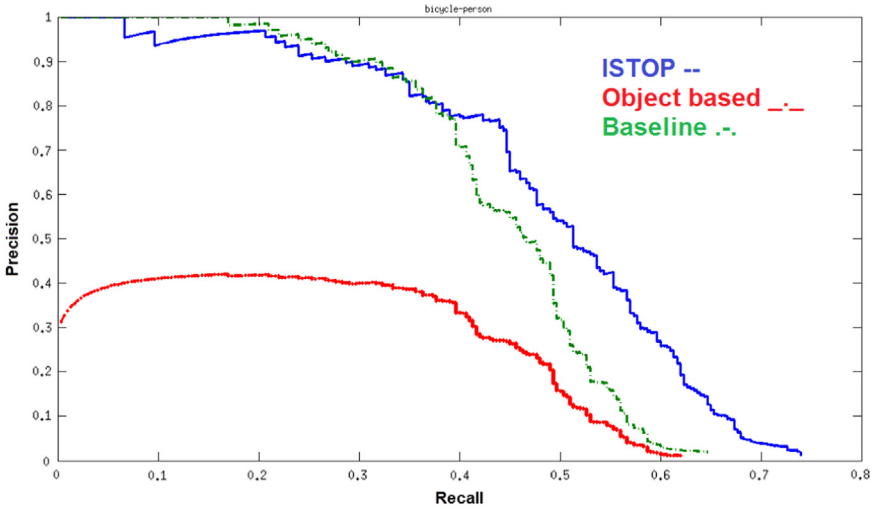

Fig. 8. Precision–Recall curves of the Object based, baseline and ISTOP model in detecting bicycle–person Visual Term.

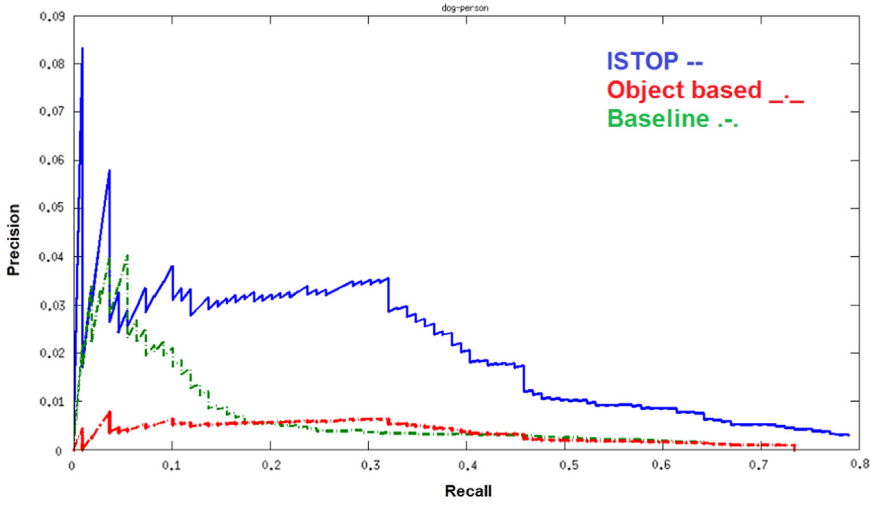

Fig. 9. Precision–Recall curves of the Object based, baseline and ISTOP model in detecting dog–person Visual Term.

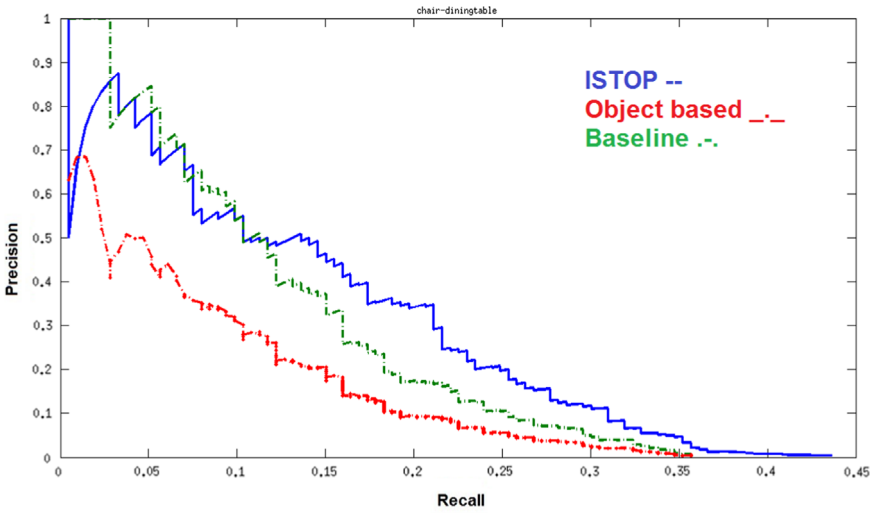

Fig. 10. Precision–Recall curves of the Object based, baseline and ISTOP model in detecting chair–dinning table Visual Term.

Table 3
Comparing the object detection by ISTOP model to DPM based on AP. Here, the objects that participate in a relation with others are mentioned.

Method	Bike	Boat	Bottle	Chair	Dining table	Dog	Horse	Mbike	Person	Sofa
ISTOP	59.7	16.0	25.6	22.3	24.7	11.2	58.6	49.6	42.8	33.4
DPM	59.5	15.2	25.5	22.4	23.3	11.1	56.8	48.7	41.9	33.6
Gain	0.2	0.8	0.1	−0.1	1.4	0.1	1.8	0.9	0.9	−0.2

approach proposed in [4] for finding co-occurrence components, in this matrix, the weight for each edge between object category A and B is computed by

$$W_{AB} = \log_{a\in A, b\in B} 2\left(\frac{(p(I|_{(a\cup b)})}{p(I|_a).p(I|_b)}\right) \quad (9)$$

where $p(I|_{(a\cup b)})$ is the probability that image I contains objects a and b that occlude each other. Similarly, the $p(I|_a)$ and $p(I|_b)$ is the probability I contains only a and I contains only b respectively.

In the matrix W, each co-occurrence category that has a strong weight (greater than zero) is candidate to compose visual terms.

4. Experiment result

In order to test the ISTOP model in extracting image concepts, we performed our model on the Pascal VOC2007 dataset [22]. Although, the used grammar in ISTOP model is top-down, the image concepts are detected in bottom-up direction as shown in block diagram in Fig. 6 and parsed samples in Fig. 7. To evaluate the model, we set up two different types of experiments and compare our model with the state of the art approaches. First, we compare on the visual terms detection to show the ability of our system to extract high level concepts. Second, we test the affect of the context and occlusion filter in object and visual term layers. Since the Pascal dataset has no labels on scenes, we could not perform our experiment on the scene level. In the future work, we may extend our experiments on the scene level by using a weakly annotated dataset including object and scene labels.

4.1. Evaluation of visual terms

Since Pascal dataset is weakly annotated, the dataset is explored to find the visual terms categories by Eq. (9). Here, 12 promising categories of co-occurrence objects are determined. To get the ground truth for visual terms, we used object labels and their bounding boxes which are available in the Pascal dataset. We explored the objects that co-occur and whose bounding boxes are near to each other or occlude each other to create labels for visual terms. By exploring the Pascal VOC 2007, we found 2538 visual terms in training images and 2181 visual terms in test images. In order to show the performance of our model in Visual Term detection, we compare it with a baseline model which is based on the visual phrase detection [3]. In addition, to show the effect of using constraints in the grammar, we also compare ISTOP with another model in which the grammar constraints are not used and is only based on the trained object filters. In this case, only the combination of detected objects is used. In Table 2, the comparison of these three approaches on the categories is demonstrated by the average precision (AP) parameter [22] and in Figs. 8–10, the precision–recall curves of three categories are shown. Here, we should mention that although our model has the ability to detect the double and multiple visual terms, in this comparison, we only consider the double visual terms. The main reason that our model performs better than the baseline approaches in detecting the co-occurrence objects is because of the individual filters to detect each object. In this way, all the possible locations are examined to find the best related position for the co-occurrence objects based on the defined constraints; while, in the baseline approach only one filter is trained for two co-occurrence objects and their relative location is considered fixed.

4.2. Effectiveness of context and occlusion filters

To show the effectiveness of context and occlusion filters, we show that the usage of context and occlusion filters can improve object detection. Since the filter templates of our model are based on the DPM [19], we show that by considering the context and occlusion filters for the co-occurring objects, the object detection becomes more accurate than before as shown in Table 3. Our model imposes constraints on co-occurring objects. Hence, the object detection of co-occurrence categories will be affected in this

Table 4
The result of using occlusion filters in visual term detection on the 12 Visual term categories based on AP.

Visual Terms	With occlusion filters	Without occlusion filters
Bicycle with Person	49.9	46.5
Boat with Person	3.4	3.2
Bottle with Dining table	4.2	4.2
Bottle with Person	17.3	17.0
Chair with Dining table	18.4	15.90
Chair with Person	8.5	4.7
Chair with Sofa	6.4	6.3
Dining table with Person	19.9	15.1
Dog with Person	2.2	2.1
Horse with Person	59.6	51.3
Motorbike with Person	35.6	29.2
Person with Sofa	13.5	12.2

case. In Table 3, the performance of our model is compared with DPM by the average precision parameter. On average our model improves object detection about 0.6 percent. This improvement is because we performed our approach on detecting occluded and co-occurring objects, but not to all objects. In our test database, the number of co-occurring objects is small relative to overall number of objects, hence the influence of our model in object detection may seem less significant but if we test our model on a dataset of occluded objects, the improvement will be significant. To show the influence of using occlusion filters, we also test the visual term detection without using occlusion filters in another experiment. As shown in Table 4, The result of this experiment proves that the usage of occlusion filters can improve visual term detection as well as object detection.

5. Conclusion

In this paper, we have proposed the ISTOP model for extracting image concepts with the ability to be trained on weakly annotated datasets. To recognize image concepts, a context sensitive grammar was proposed for parsing image from scene level to visual term, object and part level, where the context and constraints employed in the grammar impose consistency at each level. In this grammar, the Visual Term was introduced as a new concept to bridge the gap between object and scene level. The visual term represents the related co-occurrence objects as a higher concept level. The outstanding feature of the visual term is its ability to consider more than two related objects as a multiple visual term. Additionally, the co-occurrence constraints encourage compatible parts to occur together within the composition which improve object detection. The experimental results on the Pascal VOC dataset show that the ISTOP model outperforms other approaches on visual term detection, and outperforms discriminative part based model in most object detection cases.

Conflict of interest

There is no conflict of interest.

Acknowledgment

This publication was supported by the Dutch national program COMMIT.

References

[1] A.-M. Tousch, S. Herbin, J.-Y. Audibert, Semantic hierarchies for image annotation: a survey, Pattern Recognition 45 (1) (2012) 333–345.
[2] D. Zhang, M.M. Islam, G. Lu, A review on automatic image annotation techniques, Pattern Recognition 45 (1) (2012) 346–362.
[3] M.A. Sadeghi, A. Farhadi, Recognition using visual phrases, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Providence, RI, 2011, pp. 1745–1752.
[4] S.-C. Zhu, D. Mumford, A Stochastic Grammar of Images, Now Publishers Inc, 2007.
[5] N.M. Elfiky, F. Shahbaz Khan, J. Van De Weijer, J. Gonzalez, Discriminative compact pyramids for object and scene recognition, Pattern Recognition 45 (4) (2012) 1627–1636.
[6] H. Lei, K. Mei, N. Zheng, P. Dong, N. Zhou, J. Fan, Learning group-based dictionaries for discriminative image representation, Pattern Recognition 47 (2) (2014) 899–913.
[7] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[8] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, IEEE, Anchorage, AK, 2008, pp. 1–8.
[9] Y. Zhao, S.-C. Zhu, Image parsing with stochastic scene grammar, Adv. Neural Inf. Process. Syst. (2011) 73–81.
[10] F. Han, S.-C. Zhu, Bottom-up/top-down image parsing with attribute grammar, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 59–73.
[11] S. Singh, A. Gupta, A.A. Efros, Unsupervised discovery of mid-level discriminative patches, in: Computer Vision–ECCV 2012, Springer, Florence, 2012, pp. 73–86.
[12] X. Li, C.G. Snoek, M. Worring, A.W. Smeulders, Harvesting social images for bi-concept search, IEEE Trans. Multimed. 14 (4) (2012) 1091–1104.
[13] C. Desai, D. Ramanan, Detecting actions, poses, and objects with relational phraselets, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 158–172.
[14] H. Pirsiavash, D. Ramanan, Detecting activities of daily living in first-person camera views, In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Providence, RI, 2012, pp. 2847–2854.
[15] B. Yao, L. Fei-Fei, Recognizing human–object interactions in still images by modeling the mutual context of objects and human poses, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1691–1703.
[16] B. Yao, L. Fei-Fei, Modeling mutual context of object and human pose in human-object interaction activities, in: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, San Francisco, CA, 2010, pp. 17–24.
[17] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Providence, RI, 2011, pp. 1385–1392.
[18] L. Bourdev, J. Malik, Poselets: body part detectors trained using 3d human pose annotations, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, Kyoto, 2009, pp. 1365–1372.
[19] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, Discriminatively trained deformable part models, release 4, 2014, ⟨http://people.cs.uchicago.edu/pff/latent-release4/⟩.
[20] C.-N. J. Yu, T. Joachims, Learning structural svms with latent variables, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, Montreal, Quebec, Canada, 2009, pp. 1169–1176.
[21] M. Hoai, L. Torresani, F. De la Torre, C. Rother, Learning discriminative localization from weakly labeled data, Pattern Recognition 47 (3) (2014) 1523–1534.
[22] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

M.S. Zarchi is assistant professor at the Faculty of engineering, Haeri University, Meybod, Iran. He got his Ph.D. in computer engineering, Artificial Intelligence, from University of Isfahan in 2015. His main research interests include pattern recognition, image processing and image retrieval.

R.T. Tan is a computer vision scientist, currently working as a senior lecturer at SIM University (UniSIM) and an adjunct assistant professor at National University of Singapore (NUS). His main research interests are particularly in the areas of physics-based computer vision and motion analysis. Besides, he has strong interests in computer graphics and machine learning.

C. van Gemeren received his master's degree in cognitive artificial intelligence from the Humanities faculty of Utrecht University in 2012. After his graduate studies he became a Ph.D. candidate at the Science Faculty of Utrecht University. He is a computer vision researcher for the Interaction Technology group there, with a focus on the development of algorithms for interaction and mood classification in videos of groups of people.

A. Monadjemi got his Ph.D. in computer engineering, pattern recognition and image processing, from University of Bristol, Bristol, England, in 2004. He is now working as an associate professor at the Faculty of Computer Engineering, University of Isfahan. His research interests include pattern recognition, image processing, artificial neural networks, and physical demobilization and elimination of viruses.

R.C. Veltkamp is a full professor of Multimedia at Utrecht University, The Netherlands. His research interests are the analysis, recognition and retrieval of, and interaction with, music, images, and 3D objects and scenes, in particular the algorithmic and experimentation aspects. He has written over 150 refereed papers in reviewed journals and conferences, and supervised 15 Ph.D. theses. He was director of the national project GATE – Game Research for Training and Entertainment.