On the use of spatial relations between objects for image classification*

Nicolas Tsapatsoulis, Sergios Petridis, and Stavros J. Perantonis

Computational Intelligence Laboratory,
Institute of informatics and Telecommunications,
NCSR "Demokritos", Greece,
{petridis,ntsap}@iit.demokritos.gr.

Abstract. Image classification is addressed in this paper by utilizing spatial relation of detected objects in a rule-based fashion. Instances of particular object classes are detected combining bottom-up (learnable models based on simple features) and top-down information(object models consisting of primitive geometric shapes such as lines). The rule-based system acts as a model for the spatial configuration of objects, also providing a human interpretable justification of image classification. Experimental results in the athletic domain show that despite inefficiencies in object detection, spatial relations allow for efficient discrimination between visually similar images classes.

Key words: image classification, object detection, spatial realations

1 Introduction

Retrieving images based on their content is a challenging issue. Although the last decade research has being focusing on the query-by-example paradigm [1], an ambitious goal is to allow the user to formulate semantic queries through a natural language interface. Beside translating textual information into a semantically valid query, this goal also requires an association of semantic classes to their visual representations.

An approach to handle semantic queries has been to label images with coarse classes, such as indoor/outdoor and cities/landscapes, based on global characteristics of images. Such a labelling, though, tends to be inadequate in respect to realistic user-queries. At the same time, finer grain classification based directly on global image features, seems unfeasible. In more realistic scenarios, a user may wish to retrieve an image based on particular objects they appear in it. This brings up the question of detecting and classifying particular areas of images to one among a certain number of object classes. What's more, once

* This work was partially supported by the European Commission under the FP6-027538 contract.

Please use the following format when citing this chapter:

Tsapatsoulis, N., Petridis, S., Perantonis, S., J., 2007, in IFIP International Federation for Information Processing, Volume 247, Artificial Intelligence and Innovations 2007: From Theory to Applications, eds. Boukis, C., Pnevmatikakis, I.., Polymenakos, I.., (Boston: Springer), pp. 347-356.

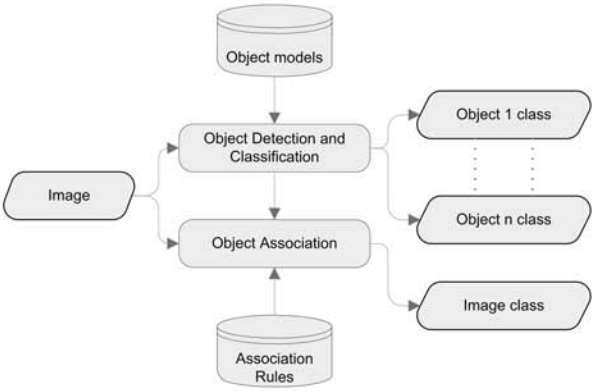

Fig. 1. Schematic diagram of the detection and classification process

this question is addressed, an overall image classification becomes conceivable, by resolving it to a particular spatial combination of objects it is made of.

In this study, we apply an object detection followed by image classification approach to detect objects and events in the athletics domain. Our object detection method results in finding image areas corresponding to a (possibly partially occluded) 2D-representation of an instance of a predefined set of object classes. To that end, we combine a top-down strategy, i.e. take into account modelling of specific object classes, with a bottom-up approach, i.e. determine region boundaries based on visual cues, as suggested in [2, 3, 4]. As a next step, we consider the image as a combination of distinct semantic objects corresponding to different area locations [5, 6, 7, 8]. We then verify the object spatial relations against a set of rules, to characterize the whole image.

In unconstrained images, there is a great variability of object classes in respect to lighting conditions and camera positions. Hence, most literature has been concentrating on very specific application domains, such as car plates recognition, horses, street scene analysis and face detection [9]. In this article, we present on-going work focused on the *athletic* domain, where (a) the objects to be identified are the humans and the athletic instruments and facilities and (b) the image is classified as a whole in respect to the athletic event it focuses on. Nevertheless, as it will be shown, our methodology allows to improve image classification results even when objects are missing or not properly detected.

2 System Overview

Semantics extraction from images has been frequently depicted as bridging the gap between concepts and their visual representations. Our approach consists of constructing this bridge with, as an intermediate abutment, the detection

of particular image areas as instances of semantic classes. The overview of our method is depicted in Figure 1.

Our assumption is that in most circumstances, even when an image can be overall described by a single word, its semantics are too complex to be detected directly by visual cues. We refer to those semantics as high-level concepts. Instead, it may be easier to decompose the image semantics into a set of interrelated concepts corresponding to distinct visual areas of the image, which may be much more easily detectable. We will refer to these concepts as mid-level concepts [10], since they serve as intermediates between visual cues and the final image classification[2]. This has the advantage of being able to explicitly supplement the extraction system with known semantics regarding the relation between the mid-level concepts and the high-level ones, thus providing useful a-priories to the extraction procedure.

To illustrate our methodology, consider an example of the athletics domain, where an image shows an athlete holding a pole and jumping over an horizontal bar, whereas a pillar is also visible. Clearly, this may be interpreted as a photo taken from a pole-vault event, as long as the relative position of these objects does indicate this. Although a direct classification of an image as a pole-vault event is theoretically possible, detecting each object separately and then associate them seems a more robust and scalable solution, if a distinction between a very visually similar event, such as high jump, is desirable.

Our methodology results in semantic labelling of images as well as of objects within images, which makes it potentially suitable for image retrieval. An important issue that arises then is how the results are further used to allow for query answering. Although early approaches employed ad-hoc methods for querying specially crafted databases [11], the approach we suggest here is to populate an ontology, which can be then further queried using a standard reasoner (see [12]). This approach has the additional advantage of using further knowledge, implied by the T-box of the ontology, to answer complex semantic queries.

3 Object Detection

Our approach to object detection is a conjunction of bottom-up and top-down techniques to detect specific objects. Namely, following a domain-independent segmentation to find a first set of segments (bottom-up approach), particular algorithms [13] [14], taking into account information regarding the colour/texture of objects, are used to detect fragments of objects classes (top-down approach). Additional information regarding the expected shape of the object classes is also used either to merge adjacent fragments of the same object class or to directly

[2] Notice that although the concepts' qualifier "mid-level" refers to their role as intermediates in bridging the semantic gap, they can also be characterised as "atomic" in that they constitute the smallest *semantic* entities detected directly through image processing techniques.

locate them in the image (top-down). In the latter case, a further combination of the segments found with the general-purpose segmentation is used to optimally adjust the object boundaries. To further distinguish among object classes in a finer grain, we extract features of the detected objects and feed them to a learnable classifier, assigning the object class with the highest score.

In the remainder of this section, we describe in details the way detection is done for three object classes: human bodies, human faces and elongated objects. The choice of these object has been such that, as it will be shown at section 4, it will enable a final image classification based on their spatial relations.

Detection of human bodies To detect human bodies, the image I is first partitioned into segments $\mathcal{S} = \{S_i\}$ using the JSEG [15] algorithm, such that

$$I = \bigcup_i S_i \tag{1}$$

To allow for more accurate object contour detection, over-segmentation is promoted, by choosing high values for the merging threshold of this algorithm. Subsequently, a small number among these segments is kept, based on whether these constitute foreground areas of the image. Foreground areas are modelled as the visually attended areas, computed with the aid of the algorithm described in [14]. The assumption here is that the human to be detected is always part of the foreground, since during photo capturing, the focus is on him. In particular, the set of segments $\mathcal{S}'$ kept as candidates for humans, comprises those having overlap precision ratio higher than a defined threshold T:

$$\mathcal{S}' = \{S_i : \frac{|M_i \cap M_F|}{|M_i|} > T\} \tag{2}$$

where M_i denotes the mask of segment S_i, M_F denotes the mask of area detected as foreground, $\cap$ denotes the logical AND operation and $|\cdot|$ denotes the sum over pixel values. A typical value for the threshold T is 0.5.

To further reduce the elements of $\mathcal{S}'$, we make use of a classifier, which decodes wether a segment is part of a body, rather than some other object class. Since the same classifier is used to discriminate among object classes, it is described separately below. Finally, adjacent partitions of this set are, then, merged and the human body is considered as the largest (with respect to area measuring) candidates after merging.

Detection of human faces To detect human faces, we rely on two essential characteristics of a face: (a) faces are skin areas having significant intensity variability due to the presence of eyes, eyebrows, mouth and nostrils and (b) faces tend to have an oval-like shape. In particular, we first detect segments containing skins, based on the combination of the JSEG segmentation algorithm with a skin-detection algorithm described in [13]. Again, over-segmentation is pursuit in order to allow discrimination between the face and neighboring naked

body parts (neck and arms)[3]. Having identified a number of potential face fragments, we proceed by selectively merging their corresponding segments. Segments are recursively merged under the condition that (a) they are adjacent and (b) the resulting segment jointly maximizes both the anticipated skin colour [13] and the *circularity* index, as compared to the largest of the component ones. The circularity index is computed as the ratio of the area of a circle having as radius the variance of the segment along its longest projection, to the actual area of the segment:

$$\frac{2\pi(\max_{|\mathbf{v}|=1} \mathrm{var}\{\mathbf{v}^\top S\})^2}{|S|} \tag{3}$$

The resulting candidate human faces segments are then given to the machine learning algorithm for a final scoring.

Detection of elongated objects Particular attention has been given to the detection of possibly occluded objects having an important elongated nature, since these are pertinent in respect to the athletics domain (horizontal bars, poles, pillars). Elongated objects have line segment characteristics and their detection involves the combination of the radon transform with hough transform. Namely, the image edges are first extracted, by using information from the gradient and the entropy of the pixels' images. Then, the matrix stemming from the radon transform, evaluated at angles with a small step (e.g. 3°), is processed by a hough transform to find optimal angles, where the intensity of accumulation is important across a wide range of pixels. Subsequently, the image-mask corresponding to each of the angles found is dilated and combined with the original image with the AND operator. The detected objects are then fed to the classifier for a final decision. To allow for discrimination among several types of elongated object classes, features such as orientation and length are also extracted.

Finer Object Classes The above methods for human face, human body and elongated objects result in image segments that possibly correspond to one of these object classes. To further enhance the ability to discriminate among these classes, as well as to discriminate among sub-classes, we make use of a classifier. This requires generating a feature vector corresponding to each image segment. The features that have been used are area, colour, area entropy (texture), circularity index, angle, and position. The generated feature vectors are then fed to a multiclass 1-vs-all extension of an RBF-SVM classifier. The class with the maximum score is then used to finally characterize the segment.

4 Ruled-Based Image Classification

In the proposed methodology for image classification, the role of rules is to provide relations between semantic entities (objects) so as to allow for an overall

[3] Notice that this is not always feasible and is actually the main reason for achieving high area recall but low area precision values (see the evaluation section below).

image interpretation. The rules are derived automatically based on the manually annotated objects and refer to spatial relations between them. Justification of using rules referring to spatial relations was, first, identified during the manual image annotation process. The question there was: "Which are the discriminatory cues that allow (humans) for identifying the class of an image given a set of available classes?". In the athletics domain it turned out that these cues were: (a) Existence of particular athletic instruments (e.g., pole, hurdles, etc), (b) posture of athlete's body, and (c) recognition of an athlete and association with her/his athletic event of expertise. The existence of particular objects, in our approach, is verified through the object detection process. As already mentioned, however, object detection (even in the context of a particular domain like athletics) is neither easy nor reliable. Thus, rules of the form "if instrument X was found then input image belongs to class Y" are error pruned. On the other hand by using spatial rules between two objects we ensure that neither object detection false alarms nor object detection misses would be able to activate a rule because a spatial relation with another object needs also to be fulfilled. As far as the athlete's body posture cue is concerned, by defining human body and human face as different objects one can define rules describing a variety of postures. Finally, the third image classification cue implies face recognition abilities so as to recognize athletes from photos. Despite the lot of work done in this area, unconstrained face recognition from images is closed to impossible

In order to construct rules concerning the spatial relations between objects we have defined a set of spatial relations that can be easily identified in the 2D-projection of a physical scene through the use of image analysis techniques. In the first stage we have used the following spatial predicates: 'is above', 'is below','is left', 'is right', 'is adjacent', 'is near', 'is above left', 'is above right', 'is below left', 'is below right'. We are currently working towards reliable automatic extraction of the 'is behind' relation.

Rule extraction Rules are automatically extracted by using the manually annotated content. Spatial relations are then computed based on the object masks. Although formal rule extraction exist [16], in a preliminary study we have constructed spatial rules by exhaustive search in our training corpus. In particular we have tried to identify rules that frequently appear in the content of a particular image class and are able to separate this image class from the other classes. A sample of derived rules are shown in Table 1. For instance, the rule with id=10 can be expressed as 'a body is below an horizontal bar'; this rule holds in the 5% (see frequency field) of training images. The 75% (see confidence field) of these images belong to the pole vault class.

Image Classification In order to classify images w.r.t a set of available classes using the above mentioned rules we use a 'rule-voting' process. That is, given the object detection results for a particular image, every activated rule votes for its class with the rule's confidence value. The overall score for a particular class is the sum of votes for this class divided by the total number of activated rules. Imagine, for example, that the rules with ids 9,10,11 hold based on the image

rule id	relation	arg. a	arg. b	frequency	class	confidence
...						
9	is above right	body	pole	7	pole vault	1.00
10	is below	body	horizontal bar	5	pole vault	0.75
11	is left	face	horizontal bar	14	high jump	0.84
12	is right	face	horizontal bar	9	high jump	0.91
...						

Table 1. Example of spatial rules

analysis results. The 'voting' score for the pole vault class is $(1+.75)/3 = 0.5833$ while the corresponding score for the high Jump class is $0.8421/3 = 0.2807$. Given that the confidence score for each rule is bounded in the $[0, 1]$ interval it is obvious that the sum of voting scores for all classes is bounded by one. However, the upper bound is rarely reached in practice. On the other hand, there are cases (images) in which no rule is activated. In this case the image class is denoted as 'unknown'. In this way images, for which the evidence for their class estimation is poor, remain unlabelled.

We should note, here, that the aim is to transfer the knowledge captured through the rule extraction process, outlined earlier, into an ontology to allow for usage of description logics. This will allow rule combination and utilization of prior knowledge already available in the ontology. A further goal is then to use the ontology to guide the object extraction process, by also detecting object's configurations unlike to appear. To give an example, in the context of pole vault and high jump images, it is unlike that a body can be above a face and both of them below an horizontal, unless a pole is also present and touches the body.

5 Evaluation Results

The performance of the presented algorithms has been evaluated based on a set of manually annotated images spatial dimensions 480×600, taken from the IAAF web site [17]. In total 140 images illustrating pole vault (69) and high jump (71) events were manually annotated by two different annotators. In order to evaluate the consistency of the manually marked areas the inter-annotator agreement (IAG), which equals the ratio of the number of pixels belonging to both annotated areas to the number of pixels belonging to at least one annotated area, was used:

$$\mathrm{IAG}_i = \frac{|M_i^1 \cap M_i^2|}{|M_i^1 \cup M_i^2|} \tag{4}$$

The ground truth area for each object instance was set as the logical OR operation between the areas marked by the two annotators under the constrained that the IAG for these annotations is higher than 0.6. In this way a ground

object class	occurrences	recall	precision	MAR	MAP	MAM
horizontal bar	111	81.1	81.8	79.0	61.8	80.1
pole	61	73.8	81.8	62.3	65.1	85.6
human face	139	62.6	64.9	91.9	48.0	66.7
human body	140	67.2	72.9	93.1	84.7	88.3

Table 2. Evaluation Results for the detection of objects. The first three columns correspond to the number of occurrences of instances of each object class, the recall and precision of the object detection method. The following three columns are percentages in respect to the object correctly identified, conveying information about the area mathching: mean area Recall (MAR), mean area precision (MAP) and Mean Area Match between annotations (MAM).

truth set was built comprising of 140 human body instances, 139 human face instances (one face was fully occluded), 111 horizontal bars instances and 61 pole instances.

Table 2 presents the results of evaluation of the object classes at image level. We consider that a segment S_i detected automatically is correct when there exist a manually annotated segment S_i^m classified under the same object class with high overlap, in the sense of eq.(4). To be fair, we consider the threshold t as a function of the manually annotated segment size, so as to be more strict (respectively less strict) for large objects (respectively small objects). To this end, we used the sigmoid-shape function

$$t(S) = a(1 + \frac{b}{1 + exp(-c|S|/|I| + 1)}) \tag{5}$$

where $|S|$ and $|I|$ are the areas of the segment and image respectively, and a, b and c are parameters set to $a = 0.1$, $b = 3$ and $c = 10$. For the segments classified as correct, the area recall and recision have been evaluated as:

$$\frac{|M \cap M^a|}{|M^a|}, \quad \frac{|M \cap M^a|}{|M|} \tag{6}$$

where M and M^a denote the mask of a detected and its corresponding manually annotated segment respectively. Their mean values across all instances of the same class is shown in Table 2. An interesting point one can notice is the poor results in face detection. This can be assigned to the variability in pose (in very few images face appears in frontal position) and a frequent partial occlusion from human body and athletic objects. The authors believe that given the difficulty of face detection in such an unconstrained environment, results are more than satisfactory. Also notice that detection of horizonal bar is more accurate than pole's, though both are detected using the same principle (elongated objects). This is due to the higher variability in shape and orientation encountered in the visual appearance of poles.

In Table 3, the evaluation results for image classification are presented. To test the generalisation performance of the rules used, we tested them on a set

Sport	Performance		Confusion Matrix		
	Recall	Precision	HighJump	Pole Vault	Unknown
High Jump	86,7%	92.2%	13	2	1
Pole Vault	75.0%	85.7%	1	12	3

Table 3. Evaluation results for image classification – Confusion matrix

object class	occurrences	recall	precision
horizontal bar	24	79.2	79.2
pole	11	63.6	70.0
human face	30	66.7	69.0
human body	32	75.0	82.8

Table 4. Evaluation Results for the detection of objects in the test set.

of 32 pole vault and high jump images not used during the rule induction and object class learning process. Object detection results for the same set are shown in Table 4. Notice that the only object class which can be used for discriminating between pole vault and high jump images is pole, since all other object classes appear in both sports. However, as can be seen from Table 4, retrieving pole vault images only upon pole existence would result in poor performance (recall 63.6%, precision 70%). Rule-based classification achieves significantly higher rates (recall 75.0%, precision 85.7%), thus alleviating false alarms and misses during pole detection.

6 Conclusion and Future work

In this paper, we proposed a methodology that allows for fine-grain image classification. At a first step, a number of key-objects with specific semantics are detected. Subsequently, the spatial configuration of these objects has been taken into account by a set of rules, to ultimately characterize the entire image. The evaluation of our approach shows that spatial relations between objects have provided substantial information for image classification. The redundancy of cues induced by both detected objects and their spatial relations allows for tempering object misses and/or misclassifications, thus rendering the overall methodology robust.

Our future plans to improve upon our methodology involve two main directions. First, we investigating one-class learning models to measure the level confidence of the objects detection. The level of confidence can then be used as a weighting factor while applying the rules. A second research direction regards rules learning, which is currently done though through exhaustive search. We expect that elaborated machine learning methods for rule extraction that, in addition, allow for complex rule formation. can further improve the accuracy and robustness of image classification.

References

1. Smith, J., Chang, S.: Visually searching the Web for content. Multimedia, IEEE **4**(3) (1997) 12–20
2. Borenstein, E., Sharon, E., Ullman, S.: Combining Top-Down and Bottom-Up Segmentation. Computer Vision and Pattern Recognition Workshop, 2004 Conference on (2004) 46–46
3. Levin, A., Weiss, Y.: Learning to Combine Bottom-Up and Top-Down Segmentation. LECTURE NOTES IN COMPUTER SCIENCE **3954** (2006) 581
4. Kapoor, A., Winn, J.: Located Hidden Random Fields: Learning Discriminative Parts for Object Detection. European Conference on Computer Vision (2006)
5. Fan, X.: Contextual disambiguation for multi-class object detection. Image Processing, 2004. ICIP'04. 2004 International Conference on **5** (2004)
6. Wolf, L., Bileschi, S.: A Critical View of Context. International Journal of Computer Vision **69**(2) (2006) 251–261
7. Amit, Y., Geman, D., Fan, X.: A coarse-to-fine strategy for multiclass shape detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on **26**(12) (2004) 1606–1621
8. Singhal, A., Luo, J., Zhu, W.: Probabilistic spatial context models for scene content understanding. Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on **1** (2003)
9. Fleuret, F., Geman, D.: Coarse-to-Fine Face Detection. International Journal of Computer Vision **41**(1) (2001) 85–107
10. Petridis, S., Tsapatsoulis, N., et al.: Methodology for Semantics Extraction from Multimedia Content. Deliverable, BOEMIE, FP6-027538 (2007) http://www.boemie.org/files/BOEMIE-d2_1-v2.pdf.
11. Li, W., Candan, K., Hirata, K., Hara, Y.: Hierarchical image modeling for object-based media retrieval. Data & Knowledge Engineering **27**(2) (1998) 139–176
12. Haarslev, V., Möller, R.: Racer: A Core Inference Engine for the Semantic Web. Proceedings of the 2nd International Workshop on Evaluation of Ontology-based Tools (2003) 27–36
13. Tsapatsoulis, N., Avrithis, Y., Kollias, S.: Facial Image Indexing in Multimedia Databases. Pattern Analysis & Applications **4**(2) (2001) 93–107
14. Tsapatsoulis, N., Pattichis, C., Kounoudes, A., Loizou, C., Constantinides, A., Taylor, J.: Visual Attention based Region of Interest Coding for Video-telephony Applications. In: Proc. CSNDSP, Patras, Greece. (2006)
15. Deng, Y., Manjunath, B.: Unsupervised segmentation of color-texture regions in images and video. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(8) (2001) 800–810
16. Bischog, W.: Learning Spatio-temporal relational structures. Applied Artificial Intelligence **15**(8) (2001) 707–722
17. IAAF: International association of athletics federations. http://www.iaaf.org (1996-2007)