DETECTING THE PRESENCE OF LARGE BUILDINGS IN NATURAL IMAGES

Jovanka Malobabić, Hervé Le Borgne, Noel Murphy, Noel O'Connor

Centre for Digital Video Processing
Dublin City University, Dublin, Ireland
jovanka@eeng.dcu.ie

ABSTRACT

This paper addresses the issue of classification of low-level features into high-level semantic concepts for the purpose of semantic annotation of consumer photographs. We adopt a multi-scale approach that relies on edge detection to extract an edge orientation-based feature description of the image, and apply an SVM learning technique to infer the presence of a dominant building object in a general purpose collection of digital photographs. The approach exploits prior knowledge on the image context through an assumption that all input images are "outdoor", i.e. *indoor/outdoor* classification (the context determination stage) has been performed. The proposed approach is validated on a diverse dataset of 1720 images and its performance compared with that of the MPEG-7 edge histogram descriptor.

1. INTRODUCTION

Semantic concepts, such as objects, people, etc. are the main instruments that humans use to navigate through and retrieve examples from large image/video databases [10]. Semantic annotation of large image/video databases is thus essential if ease of access and use is to be ensured. Inferring the presence or absence of high-level semantic concepts from low-level visual features is a research topic which has attracted a considerable amount of interest lately. Our objective in this paper is to detect the presence of a large *building* object (i.e. *outdoor architecture* according to [10]) in an outdoor colour image in a general purpose collection of digital photos taken by a ground-level camera in an otherwise unconstrained environment. In the image of interest, a *building* is either a single dominant object or one of the dominant objects. We aim to show that the feature representation based on a few carefully selected and physically meaningful low-level features, coupled with the high generalisation ability of the SVM classifier engine, may be sufficient to detect some high-level concepts, such as buildings. As there exists a number of methods that address the issue of *indoor/outdoor* classification of consumer photographs [9][12], we assume the implicit presence of contextual information in the form of an *indoor/outdoor* label.

The paper is organised as follows: we start with an overview of related work in Section 2. In Section 3, we present our approach, describe the extraction of low-level edge orientation features and follow a brief overview of the Support Vector Machines (SVM) classifier. We conclude with performance evaluation and discussion of our experimental results and comparison with similar work in Section 4.

2. RELATED WORK

A significant portion of research work in the area of building detection focuses on building detection in a constrained environment using multiple images of a scene (e.g. building detection in aerial photography). The majority of researchers, addressing either aerial or ground-level photography, utilise some sort of edge distribution-based feature as a low-level descriptor.

Vailaya *et al.* [13] developed a procedure to qualitatively measure the saliency of a feature towards a particular classification problem based on the plot of the intra-class and inter-class distance distributions of that feature. They show that a specific high-level classification problem can be solved using relatively simple low-level features geared for the particular classes. The edge direction coherence histogram was found to have sufficient discrimination power to distinguish between cityscape and landscape images (an edge pixel is considered *coherent* if it belongs to a connected component in a given direction whose size is at least 0.1% of the image size). This feature is geared towards discriminating structured edges from arbitrary edge distributions. The presence of human-made objects or structure in an image results in an edge direction histogram that exhibits peaks at/around the significant edge directions, whereas the edge distribution for *nature* images appears to be of random nature, i.e. distribution usually appears to be flat.

The Dorado and Izquiredo [2] approach is based on the MPEG-7 edge histogram descriptor (80-point histogram representing local distributions of directional edges within an image: 0°, 45°, 90°, 135°, and non-directional) and on local and global distribution of edges. The approach exploits rough matching and problem domain knowledge through user relevance feedback while classification is performed based on rule-based fuzzy inference. The image is spatially divided in 16 equally sized sub-images, each of which is further divided into a given number of non-overlapping small square blocks – the block size depending on the sub-image size. The blocks are divided into 4 sub-blocks and passed through 5 filters to assign them to a corresponding edge category. The number of blocks per edge category is counted to compute the edge distribution within a sub-image and 80-bins (16 sub-images x 5 bins each) summarise the distribution of each edge category. Fine-tuning is performed through relevance feedback.

The approach of Iqbal and Aggaraval [3][4] for detection of large man-made objects, such as buildings, bridges, towers, etc., is based on perceptual grouping of image primitives according to Gestalt principles of perceptual grouping (continuity, closure, proximity, co-linearity, co-circularity, symmetry, parallelism). Lower-level primitive image features, such as line/edge segments, are grouped hierarchically into higher-level structures aiming to reach a meaningful semantic structure. The goal of grouping is to identify image features that are likely to have arisen from some scene properties rather than accidental arrangements ("the principle of non-accidentalness"). For building images, a 3-component feature vector is used to represent an image to be classified into 3 classes: *building, intermediate* and *non-building.* Features used are: number of "L" junctions, "U" junctions and "significant" parallel lines in the total number of "retained" lines. In [5], they combine features (based on perceptual grouping), colour features and texture features into a 66-dimensional feature vector to represent an image. Their experiments confirm the intuitive expectation that colour information does not have sufficient discriminative power for *building/non-building* classification. Their method achieves good classification performance for broader classes such as man-made structures, but performs modestly on subclass classification within the man-made class.

Common to all three approaches outlined above is the focus on edge/line segments features and the use of orderliness/regularities that the presence of human-made objects in a scene generates in terms of edge distribution.

In the work described here, we approach the problem of *building/non-building* classification of the whole image using simple low-level features suited for the classification problem at hand, resulting in a low-dimensional feature space. Our approach for detecting the presence of large buildings in consumer photographs is based on multi-scale analysis, from global to local level and it relies on explicit edge detection. An SVM classifier engine is employed to infer the information about the presence of a large/dominant *building* object from the edge orientation-based features. We show that a few simple features with physical meaning coupled with the high generalization ability of the SVM can yield decent classification performance comparable to that of the existing approaches. The key aspects of our approach are low-dimensionality and simplicity.

3. PROPOSED APPROACH

3.1. Motivation

Our objective is to detect the presence of a large *building* object in an outdoor colour image in a general purpose collection of digital photos taken by a ground-level camera, at a close or medium distance, in an otherwise unconstrained environment. In the image of interest, a *building* is either a single dominant object or one of few dominant objects in a possibly cluttered scene, with a complex background with frequently occurring occlusions.

Our approach is based on classification of low-level feature representation of an entire image and a simple observation: the most commonly occurring views of a *building* in a standard (i.e. non-artistic, general purpose consumer) photo can be summarised into six main types as shown in Fig 1.

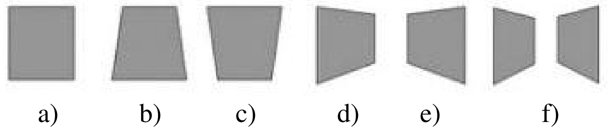

Fig. 1 A building projection as a function of common viewing angles: a) frontal view, b) frog's view, c) bird's eye view, d) view from right, e) view from left, f) "street"

The presence of a dominant human-made object in a scene generates strong evidence in the form of straight or elliptical line segments and edges [3][4][5]. Given that there is huge variation within the *building* class in terms of possible shapes that different types of buildings may take, as illustrated by Fig. 2, we take the view that a coarse modelling of building shape/geometric properties is an appropriate approach. Dominant edge orientations of *building* object boundary edges and edges due to windows, doors, etc., are in most cases a combination of near-vertical and near-horizontal with near-45°, or near -135° degrees. Examination of the 36-bin edge orientation histograms of *building, nature,*and *structure* images in Fig.3 shows that "interesting events", which distinguish between *building* and *non-building* images, (e.g. large peaks), occur at around angles such as 0°, 45°, 90°, 135° depending on the viewing angle. Fig. 4 illustrates the contributions of each of the significant edge orientation intervals to the total edge magnitude image. This indicates that it may be sufficient to base our representation on relevant subsets of the edge histogram instead of the entire histogram. The edge segments are, in accordance with the Gestalt principles, expected to obey the rules of good continuity and colinearity.

It is assumed that concepts which are large are also semantically important and that they usually, in consumer photographs at least, occur around the centre of the scene. We therefore incorporate localised information based on analysis of the central rectangular region 25% of the image size.

3.2. Low-level feature extraction – edge orientation features

The edge direction histogram is a global shape descriptor. It captures the general shape information in the image and it has been shown to be suited for use in a general purpose database. The fact that it that does not require segmentation as a prerequisite is a significant advantage considering that object segmentation is still a difficult problem. Other advantages of edge histogram include its invariance to translation in the image and robustness to partial occlusion. However, edge histogram features are inherently neither scale nor rotation invariant. Scale invariance, which in this context means invariance to the absolute size of the object, is achieved by normalising the histogram by sum of weighted contributions of all edge pixels considered. In this way we are able to deal with images (and buildings) of different sizes, avoiding the need for preprocessing.

The use of edge orientation histograms instead of edge direction histograms allows us to effectively reduce the number of bins considered, while retaining the relevant information (e.g. on parallelisms) by reinforcing the relevant peaks in the 0 to 180 degrees range.

Fig. 2 Variety of building shapes

Detection of certain features in an image is optimal at a certain scale and the correct scale or the appropriate size of the neighbourhood depends on the scale of the object under investigation. The exact size of t*he object is generally not known a priori,* thus optimal processing of an image requires the representation of an image at different scales [6][12]. As the appropriate scale is unknown (it is only known that a building is at a close or a medium distance from camera), we adopt a multi-scale approach to edge detection and apply a Canny edge detector at three scales. Scaling is achieved by smoothing with Gaussian kernel with values of σ = 1; 1.5; and 2 empirically selected. The thresholds for hysteresis thresholding were set to 0.3 and 0.9 so as to ensure that most of the edge evidence generated by the texture edges is discarded while that due to edges corresponding to boundaries is retained. Non-maximum supression ensures that all edges are one pixel wide.

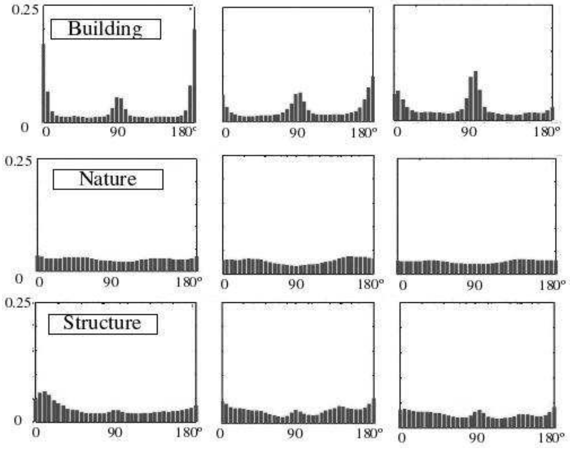

Fig. 3 Comparison of normalised 36-bin edge orientation histograms for *building*, *nature* and *structure* images

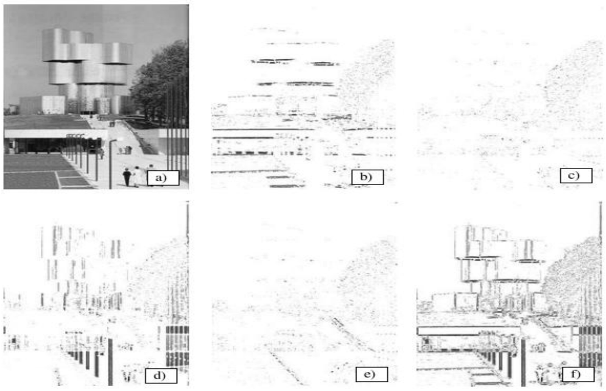

Fig. 4 Edge magnitude components pertaining to significant orientation intervals: a) original image, b) near horizontal, c) near-45, d) near vertical, e) near-135, and f) all four relevant edge orientation intervals.

In addition to global edge detection we extend our search for evidence to a subblock corresponding to the central 25% of the image as we assume that if a building is really a dominant object there must be strong evidence of human-made structure in the centre of the image. We construct a five-bin histogram at each scale, globally and locally: four bins correspond to the following edge orientation intervals: F_1=[0,10]+[170,180], F_2=[35,55], F_3=[80,100], F_4=[125,145], and one bin is used for non-relevant edge pixels (i.e. all other edge pixels). Edge pixels contributing to the first four bins are referred to as "relevant" in the following. Each 5-bin histogram is then normalised by the sum of all five bins. The 24--dimensional feature vector is then formed by discarding the fifth bin and by concatenating the remaining 4 bins for each of 2 zones at each of 3 scales.

Three versions of the approach, using different weighting schemes, are implemented. We compute the 5-bin histograms, one for each region at each scale as follows:

$$H_{ej}(i) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} W_{ei} \, I_{ei}(m,n) \, I^{j}_{reg}(m,n)$$

$$i=1,2,3,4,5; \quad j=1,2$$

where $H_{ej}\ (i)$ is an edge histogram bin corresponding to orientation i and region j (the first region is the entire image and the second region is the central 25% of the image). W_{ei} is the weight assigned to the contribution of an edge pixel with orientation i, $I_{ei}\ (m,n)$ is an edge image component for orientation i, and $I^{j}_{reg.}\ (m,n)$ is a binary zone image (with value 1 for pixels in the region of interest, value 0 elsewhere).

In the first version, the edge pixel contribution to a given bin is weighted by the gradient magnitude, and the five-bin histogram is normalised by the sum of all edge pixel contributions in the image region being analysed so as to account for different image sizes. In the second version a weighting scheme which favours contribution of edge pixels more likely to belong to linear lines is introduced. The idea is to increase the importance of the relative contribution of the pixels that obey the good continuity rule. As illustrated in Fig. 5, the 8-neighbourhood is examined for edge pixels with the same quantised orientation, termed *coherent* pixels and the highest weight W_{ei}=1.3 is assigned to an edge pixel contribution both of whose neighbours lie in direction perpendicular to its gradient direction (in case of one such neighbour weight W_{ei}=1.2 is assigned, in case of two weight W_{ei}=1.3 is assigned). In the third version, a stronger weighting is used and the weights for *coherent* pixel contribution are increased to W_{ei}=2 and 3 respectively.

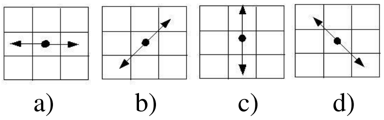

Fig. 5 Coherency check in 8-neighbourhood for the edge angle θ of the central pixel: a) $\theta \in [0,10] \cup [170,180]$, b) $\theta \in [35,55]$, c) $\theta \in [80,100]$, d) $\theta \in [125, 145]$

3.3. Low-level feature classification

The Support Vector Machine (SVM) [1] is a popular learning algorithm which has been extensively used in a number of applications, such as text classification, feature extraction and hand-written digit recognition [8]. The SVM is characterised by high generalisation ability and is based on the idea of finding the hyperplane that best separates two classes after mapping the training data into a higher-dimensional feature space via some kernel function Φ. The SVM classifiers are based on the hyperplanes of the class:

$$(\boldsymbol{w}.\boldsymbol{x})+b=0\,, \quad \boldsymbol{w} \in \mathbb{R}^N,\ b \in \mathbb{R}$$

where $\boldsymbol{w}$ is a weight vector, $\boldsymbol{x}$ is the training data, b is a threshold. The corresponding decision function $f:\mathbb{R}^N \rightarrow \{\pm 1\}$ is:

$$f(x)=sign((\boldsymbol{w}.\boldsymbol{x}) + b)$$

where $\boldsymbol{x}$ is a feature vector to be classified. The hyperplane is constructed by solving a constrained optimisation problem whose solution, a weights vector $\boldsymbol{w}$, is expressed in terms of a subset of training examples that lie on the margin: $\boldsymbol{w}=\sum_i \alpha_i \boldsymbol{x}_i$. This subset of training examples, called Support Vectors, carries all the relevant information contained in the training set. Thus the final decision function, $f(x)=sign(\sum_i \alpha_i (\boldsymbol{x}.\boldsymbol{x}_i)+b)$, where $\boldsymbol{x}$ is a new feature vector to be classified and $\boldsymbol{x_i}$ are support vectors, depends only on the dot product of the feature vectors. One of the advantages of SVM over other classifiers is its speed, as the number of points that the SVM evaluates when a new point is classified is equal to the number of support vectors (usually significantly smaller than the number of training examples). We use the SVM^{light} [7][8] classifier which outputs a confidence measure for each test sample: the sign of which determines the class membership (if the score is positive, the example is labelled as a class member and a non-class member in case of negative score) while its absolute value gives an indication of the classification decision confidence i.e. the distance from the separating hyperplane.

4. PERFORMANCE EVALUATION

4.1. Dataset

In order to evaluate the performance of the method we use a diverse database of 1720 images (consumer photographs), split into two sets: 2 different subsets of 200 images were used for classifier training/learning and the remaining 1520 images were used to evaluate the performance of the trained classifier. The dataset consists of images of arbitrary sizes in both portrait and landscape format. The images were collected from various sources: photo albums on the Internet, scanned from personal photographs and donated digital photographs. *Non-building* images include several sub-classes: nature (beaches, forest, field, water body, sunset, sunrise, etc.), large human-made-structure-other-than-building (boats, ships, cars, wheels, monuments, windmills, etc.), close-ups of flowers and fruit, animals and people (close-ups and medium distance).

Particular care was taken to ensure that the data set is almost evenly split between *building* (769 images) and *non-building* (751) images, that the dataset includes images of objects that may easily be misclassified as buildings (113 *structure* images or 15% of non-building images) and that intraclass variance of the *building* images is sufficiently large (churches, cottages, skyscrapers, castles, huts, family houses, etc).

For the creation of a ground truth we apply a single label model assuming that all images can be singly labelled. Each image was labeled by two human subjects and a class was assigned based on the subjects' perception of the dominant class in a given image.

4.2. Classifier training

Leave-one-out validation on the training set of 200 images (100 building, 100 non-building) is performed in order to determine the classifier parameters. The SVM with linear kernel is trained with different values of cost factor (which controls the ratio of misclassification penalty for the class and non-class members and corresponds to translation of the separation plane). As a criteria for selection of the SVM model we use the break-even-point on the training set (value for which recall and precision on the training are equal) and a classifier with cost factor of 1.3 was selected.

4.3. Classification based on low-level features and discussion of experimental results

As a performance measure we use classification accuracy, recall and precision on the test set of 1520 images. Classification accuracy is a fraction of all images which has been assigned to a correct class. Recall is a fraction of *building* images which has been assigned to a *building* class whereas the precision is a fraction of images assigned to *building* class that actually belong to a *building* class.

4.3.1. Experiment 1- the effect of coherency weighting

In order to determine the impact of weighting, we compare the performance of three different versions of the method: with edge magnitude weighting, weak coherency weighting and strong coherency weighting with the MPEG-7 edge histogram descriptor. The results presented in Table 1 show that the strong coherency weighting scheme outperforms both weak coherency weighting and edge magnitude weighting, as well as the MPEG-7 edge histogram descriptor.

	Accuracy	*Recall*	*Precision*
Grad. Magnitude Weighting	85.52	81.27	89.16
Coherency Weak Weighting	87.30	83.38	90.81
Coherency Strong Weighting	**88.22**	**84.01**	**92.02**
MPEG-7 Edge Hist. Descript.	84.93	79.45	89.59

Table 1. Comparison of experimental results for different methods (200 training images, 1520 test images)

4.3.2. Experiment 2 – the effect of local information

In order to verify the hypothesis that the inclusion of the localised edge information pertaining to the central 25% of the image actually improves classification performance, we compare the performance of the 12-component global and local feature representations with the 24-component feature representation (global+local information) for strong coherency weighting. The results in Table 2 confirm that, for this particular dataset at least, the incorporation of localised information positively affects the classification rate.

	Accuracy	*Recall*	*Precision*
12-component (local)	87.67	82.74	92.06
12-component (global)	86.18	81.40	89.96
24-component (global+local).	**88.22**	**84.01**	**92.02**

Table 2. Comparison of performance of 12-component and 24-component representation for strong coherent weighting (200 training, 1520 test images)

The examination of misclassified images in both cases shows that this improvement is due to a reduction in the misclassification of *structure* images.

By closely examining the misclassified images we observe that most frequently misclassification occurs in the case of scenes containing dominant human-made structures other than buildings with edge distributions similar to that of buildings such as those shown in the top rows of Fig 6. In other cases, the misclassification occurs due to strong regular textures such as the presence of tree trunks in a close proximity to camera as can be see in Figure 6 (a forest in the bottom row).

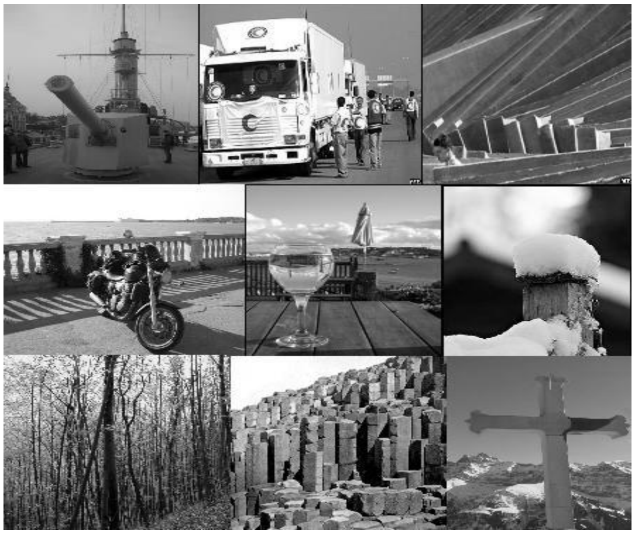

Fig. 6 Typical non-building images misclassified as buildings

The another difficult example is the Giant's Causeway (a naturally occurring outcrop of hexagonal basalt columns in Northern Ireland shown in Fig. 6 in the middle of the bottom row,) that exhibits exceptionally high degree of regularity and features we normally associate with human-made objects. We also observe misclassification of *building* images due to the fact that edge orientation based features are not rotation invariant, as can be seen in Fig. 7 (the two building images on the left were misclassified with high degree of confidence).

Performance of our approach is comparable to that of the existing approaches. However, we have to emphasise that we used our own dataset and a different number of training examples so that we are not in a position to make a fair comparison. Dorado *et al.* report similar recall and precision on a test set of 3000 TREC images using 115 images for training. The user interaction improves the recall and precision to 86.31 % and 86.25% respectively. Iqbal and Aggarwal validate their approach on 120 images (using 30 images for training) and report the recall of 80% and precision of 83.72%. We compare with the performance of a standard MPEG-7 edge histogram descriptor and as can be seen from Table 1, our approach outperforms SVM classification based on an MPEG-7 edge histogram descriptor on a common dataset.

5. CONCLUSIONS

In this paper, we presented an approach to *building/non-building* classification of outdoor consumer photographs based on a few simple edge-orientation features with physical meaning, extracted at three scales, and used in conjunction with an SVM classifier engine. Experimental results on a diverse dataset of 1720 images show that the performance of our method is comparable to that of the existing approaches. However, the results also show that an improvement is required in order to overcome the lack of rotation invariance and reduce misclassification between buildings and other human-made structures. Future work will include extensive comparison with other techniques.

6. ACKNOWLEDGEMENTS

This material is based on works supported by Enterprise Ireland under Grant No.03/IN.3/I361 and the Ulysse research project ReSEND (FR/2005/56). The support of IST-2000-32795 SCHEMA, RINCE and HEA National Development Plan is gratefully acknowledged. This work is partly supported by the European Commission under

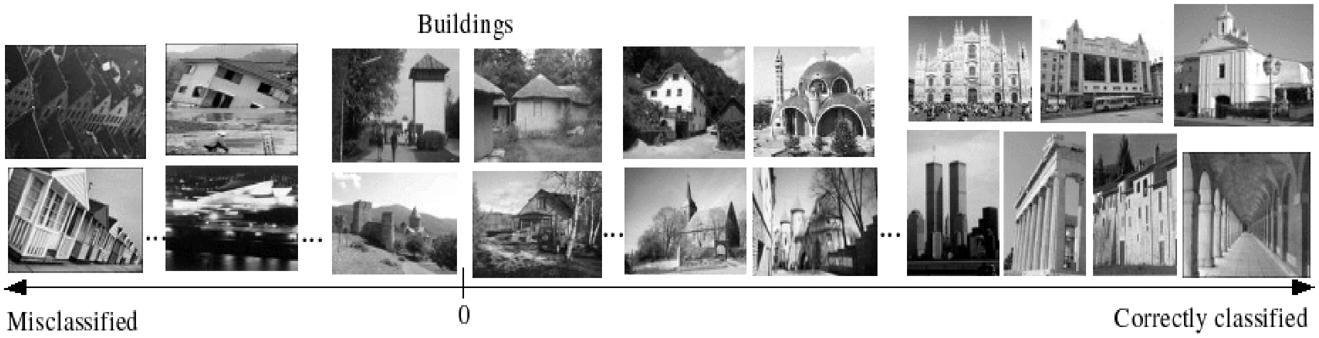

Fig. 7 Classification result for *building* images in order of decision confidence

contract FP6-001765 aceMedia (URL: http:// www.acemedia.org). In addition we gratefully acknowledge Patrick Gros of INRIA Rennes, for his comments and access to the image database, and Augustin Gavrilescu of Vrije University Brussels, for supplying part of the image corpus.

7. REFERENCES

[1] C.J.C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition”, Data Mining and Knowledge Discovery, 2, pg 121-167, 1998.

[2] A. Dorado and E. Izquiredo, “Exploiting Problem Domain Knowledge for Accurate Building Image Classification”, *Proceedings of CIVR 04*, Dublin, Ireland, July 2004.

[3] Q. Iqbal and J.K. Aggarwal, “Applying Perceptual Grouping to Content-based Image Retrieval: Building Images”, *Proc of the IEEE Int Conf on Computer Vision and Pattern Recognition*, Fort Collins, USA, June 1999

[4] Q. Iqbal and J.K. Aggarwal, “Lower-level and Higher-level Approaches to Content-based Image Retrieval”, *Proc. of the IEEE South West Symposium on Image Analysis and Interpretation*,Austin, Texas, USA, pp. 197-201, April 2000.

[5] Q. Iqbal and J.K. Aggarwal, “Combining Structure, Colour and Texture for Image Retrieval: A Performance Evaluation”, *Proc of Int. Conf. on Pattern Recognition (ICPR),* Quebec, Canada, 2002

[6] B. Jahne, *Digital Image Processing, Concepts, Algorithms and Scientific Applications*, 4th edition, Springer-Verlag Berlin Heidelberg, 1997

[7] T. Joachims, SVMlight,http://svmlight.joachims.org/

[8] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features”, *Proc. 10th Eur. Conf. on Machine Learning,* pp. 137–142, 1998

[9] J. Luo and A. Savakis, “Indoor vs Outdoor Classification of Consumer Photographs using Low-level and Semantic Features”, *Proc.of IEEE Int. Conf. On Image Processing, ICIP 2001*, Thessalonki, Greece, Oct 2001.

[10] A. Mojsilović and B. Rogowitz, “Capturing Image Semantics with Low-Level Descriptors”, *Proc.of IEEE Int. Conf. On Image Processing, ICIP 2001*, Thessalonki, Greece, Oct 2001.

[11] M. Sonka, V. Hlavač, and R. Boyle, *Image Processing, Analysis and Machine Vision*, 2nd edition, Brooks/Cole Publishing Company, 1999

[12] M. Szummer and R.W. Picard, “Indoor/outdoor Image Classification”, *IEEE Intl Workshop on Content-based Access to Image and Video Databases*, January 1998.

[13] A. Vailaya, A. Jain and H.J. Zhang, “On Image Classification: City Images vs. Landscapes”, *Pattern Recognition*, 31:1921-1936, 1998