Analysing Image-Text Relations for Semantic Media Adaptation and Personalisation

Mark Hughes[1], Andrew Salway[2], Gareth Jones[1] and Noel O'Connor[1]
[1]Centre for Digital Video Processing, Dublin City University, Ireland
[2]Burton Bradstock Research Labs, UK
{mhughes, gareth.jones}@computing.dcu.ie, andrew@bbrel.co.uk, oconnorn@eeng.dcu.ie

Abstract

Progress in semantic media adaptation and personalisation requires that we know more about how different media types, such as texts and images, work together in multimedia communication. To this end, we present our ongoing investigation into image-text relations. Our idea is that the ways in which the meanings of images and texts relate in multimodal documents, such as web pages, can be classified on the basis of low-level media features and that this classification should be an early processing step in systems targeting semantic multimedia analysis. In this paper we present the first empirical evidence that humans can predict the main theme of a text from an accompanying image, and that this prediction can be emulated by a machine via analysis of low-level image features. We close by discussing how these findings could impact on applications for news adaptation and personalisation, and how they may generalise to other kinds of multimodal documents and to applications for semantic media retrieval, browsing, adaptation and creation.

1. Introduction

In the field of semantic media analysis very little is known about how the meanings of different media types combine in multimodal documents. This fact creates a severe limit on the automatic analysis of multimedia data and on dependent applications for semantic media adaptation and personalisation. In [1] a variety of image-text relations were postulated in an attempt to account for the different ways in which the meanings of images and texts can combine in multimodal documents such as web pages and hypermedia presentations. It was suggested that such image-text relations could be recognised by humans, and potentially by machines, on the basis of low-level image, text and page layout features, but this was not established empirically. It was also suggested that image modality, on a scale from realistic-abstract, or photographic-graphic, was a cue to whether an image depicts the specific or general person. It was proposed that an image depicting a specific person has a realistic modality, which is realised by sharp focus, deep colour and high brightness. In [2] it was argued that the automatic classification of image-text relations as an early step in semantic media analysis would enhance the integration and fusion of multimedia data in applications for semantic retrieval, browsing, adaptation and creation.

In our ongoing work we are investigating image-text relations in online news stories which all comprise text and an associated image – typically a photograph. Firstly, we are interested to find out more about how humans read these multimodal documents, in particular how seeing the image influences their expectations of the text, and vice versa. Secondly, we are aiming to classify image-text relations automatically so that predictions of how the meanings of texts and images are related can be factored into semantic media adaptation and personalisation.

Section 2 reports an experiment to test the hypothesis that humans can predict the main theme of a text by looking quickly at an associated image. We found that by seeing pictures of people that accompany 80 online news stories, 25 subjects could predict very accurately whether the story was about the specific person/people depicted in the image, or about a more general theme. The positive findings from this experiment encouraged us to look into low-level features that could be used to make this prediction automatically. Using a face detection algorithm set to detect large full-frontal faces, a measure of variation in image sharpness across the image and certain features intended to correlate to image modality, we are able to correctly classify photographs into Specific or General categories in 82.5% of 80 online news stories – see Section 3. In Section 4, we discuss the potential impact of these findings on applications for news adaptation and personalisation, and consider the more widespread applicability of knowledge about image-text relations for semantic media analysis and the creation of multimedia information.

2. Human Classification of Image-Text Relations

The aim of this experiment was to test the hypothesis that low-level image properties can enable humans to predict something about the meaning of the text associated with an image. Two sets of 40 online news stories were gathered from news.bbc.co.uk, www.guardian.co.uk, www.cnn.com and www.thesun.co.uk. All collated web pages comprised the main text of the news story and an accompanying photograph of one or more people. In one set, all the photographs showed the specific person that the story was about: in the other set the person was unnamed in the story which was about some general theme. We determined the *Specific* vs *General* distinction by reading the news stories – in most cases it was enough to read the first few lines. The page layout and relative size of image and text did not vary between Specific/General, though they did vary between news websites.

The web pages were prepared so that the text was blurred to make it unreadable, but so that it was still obviously a web page with only the image clearly visible. The 80 modified web pages were then shown to 25 subjects for about 3s each and the subjects were asked to decide for each page whether the image was Specific or General, i.e. was the story about the specific person shown in the image or about a general theme. The subjects were shown 2 examples of each category before the experiment started, see Figure 1 for an example of each.

For 73 out of 80 online news stories (91%), 21 or more of the 25 subjects gave the correct classification of Specific or General based on seeing the modified web page with only the image visible clearly; more results are given in Table 1. Looking through the stories that were correctly classified nearly unanimously, we came up with two sets of reasons which might explain why humans can do this task so quickly and reliably.

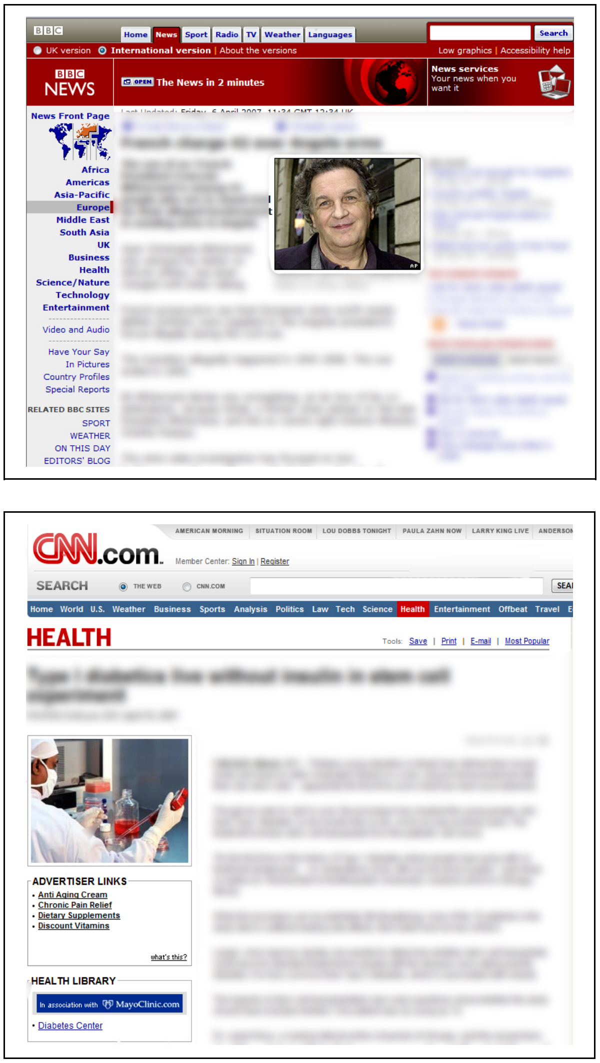


Figure 1. *An example of a story about the Specific person depicted in the photograph (top) and a story about a General theme (bottom)*

Table 1. *Results for the task of determining Specific or General*

Number of Subjects giving correct response	*Number of online news stories*
25/25	27/80 (34%)
>24/25	51/80 (64%)
>23/25	64/80 (80%)
>22/25	71/80 (89%)
>21/25	73/80 (91%)

One set of reasons is to do with subjects' knowledge about people in the media and expectations about particular kinds of news stories. Some people are recognisably famous so the story is likely to about them specifically; also distinctive are criminals' mug shots that could cue subjects to a Specific classification. In the case of some General classifications it could be that subjects are used to seeing pictures of unnamed soldiers, police, protesters, etc. accompanying stories about war, accidents, demonstrations, etc. Of course these kinds of inferences would be non-computable given current limits on computer vision and artificial intelligence.

However, through manual inspection of the online news stories we noted some characteristics of Specific and General images that could be computed. In many Specific cases the photographs show people with their whole face visible, and often the people are looking directly at the camera. Furthermore, the face is relatively large and centered compared to General cases. Finally, we noted that in Specific cases the face tends to be in sharper focus than the rest of the image – whereas the sharpness seems more constant across General images.

Interestingly, in each of the 7 stories that 5 or more subjects classified incorrectly, one or more of the 'rules' noted above was broken. In a General example – a story about Iran's nuclear ambitions – the Iranian president is potentially recognisable, although his face is small and not full-frontal: it was incorrectly judged Specific by 10/25 subjects. In another General example – a story about the war in Iraq – an unnamed soldier is photographed looking straight to camera, his face quite large and in sharp focus: 5/25 subjects incorrectly judged it Specific. In a Specific example – a story about a woman and child left stranded by a vehicle recovery service - the people that the story is about are photographed such that their faces are relatively large and full-frontal, but they are not centered in the photograph and are not looking at the camera: it was incorrectly judged General by 6/25 subjects. See the Appendix for these three examples.

We conclude that it is possible for humans to predict something about the meaning of text associated with an image on the basis of low-level image features, though we were not able to factor out the effects of subjects' world/media knowledge. It seems that there are some conventions in online news production that guide the selection and editing of photographs to accompany stories, and that these conventions manifest in low-level image features that could be used to automatically classify image-text relations.

3. Automatic Classification of Image-Text Relations

3.1 Choice of Features

In this section we report how we tried to emulate the ability to classify Specific vs. General images automatically based on the extraction of low-level image features only. Our choice of features is based in part on the 'rules' identified in Section 2, i.e. that in Specific examples the accompanying photographs: (i) show larger full-frontal faces, and (ii) the face is sharper in contrast to rest of the image, i.e. the variation in sharpness across the image is higher than in General examples. We also consider some features identified in [1] to represent a realistic image modality, which could help to differentiate between the classes.

3.2 Description

The ‘rules’ identified in Section 2 tell us that images that were classified as Specific seemed to have large full frontal faces (full-frontal defined as both eyes, nose and mouth of face are visible within image) usually located near the centre of the image. In contrast, in images classified as General, if visible faces were present then they were not full frontal and were quite small and less often centred. Therefore we decided that it was necessary to detect only large faces, which were full-frontal to help differentiate between the Specific and General classifications of images. We used an appearance-based face detection algorithm as described in [3] to detect faces within our images. This method of face detection will only detect faces which are full-frontal and which are quite large (larger than 16 x 16 pixels).

In fact, after testing the discrimination values of each feature we discovered that the facial features discriminated the best therefore we decided to use two facial features in our feature vector. We used the relative position of faces within an image and the number of faces detected in an image. To obtain the relative position of faces in the image, the normalised distance from the center of each detected face to the center of the image was calculated. If there is more than one face detected in an image then the average distance is calculated.

We implemented a metric termed 'Variation in Sharpness' that was intended to capture the difference between photographs in which a face is sharply focused against a fuzzier background, and photographs with a relatively constant level of sharpness across the image. To calculate the sharpness variance feature we use a technique similar to [4]. First we perform edge detection on the image using the horizontal Sobel operator. It has been shown in [4] that only the horizontal Sobel operator is necessary to calculate a sharpness measure of an image. We then split the image into 8x8 image blocks. Each image block is examined and the average edge width within each block is calculated. The variance of these average edge widths is then calculated and this gives us our sharpness variance value for the image:

$$\sum_{i=1}^{n}(x_i - \mu)^2$$

where n is the number of image blocks, x is the average edge width of an image block and μ is the mean edge width.

We decided to also extract images features which could represent realistic modality in an image. Motivated by [1] we chose the following three features to represent this.

Average Intensity. This was intended to correlate with the perception of image brightness. To calculate the average intensity of an image, every pixel in the image is converted from the RGB to the YUV colour space. The average Y (luminance) value of every pixel in an image is then calculated to give the overall average intensity.

Colour Variance. This was intended to correlate with the perception colour richness. Since only the colour variance among the dominant colours in an image was desired, the colour space is divided into eight bins: black, white, red, green, blue, yellow, cyan and magenta. Each pixel value is examined and stored in its appropriate bin using the smallest Euclidean distance between the respective colour values. The number of pixels in each bin is examined and compared against a threshold. The variance of the colour values contained in the bins that passed the threshold is calculated.

Global Sharpness. This was intended to correlate with the perception of how sharply focused the image is. For this we wanted to measure the sharpness based only on sections of an image that were in focus. The sharpness measure outlined in [4] and used above for our sharpness variation measure was used again here. Edge detection is first performed on the image using the Sobel operator. In this case, each image block above a certain threshold is marked as an edge block. The average edge width is then calculated across all these edge blocks to give the overall sharpness measure.

Our complete feature set thus consisted of:

1) Number of faces within image
2) Relative position of faces within an image
3) Variation of sharpness
4) Average Intensity
5) Colour Variance
6) Global Sharpness Measure

3.3 Classification results

We tested 2 commonly used types of computational classifiers corresponding to: 1). K-Nearest Neighbour classifier and 2). Support Vector Machine. We used a training set of 200 images (100 General, 100 Specific) to train our classifiers. These training images were gathered from the same news websites as the test image set and were manually classified as belonging to either Specific or General classes.

3.3.1 Support Vector Machine

A support vector machine (SVM) is a popular supervised learning method for classification [5]. The support vector machine implementation that we used is called SVMLight [6]. SVMLight is a highly configurable support vector machine implementation. The feature values extracted from the images were converted to a format that is compatible with SVMLight. The features were then normalized ensuring each feature value lies in the range [0,1]. An SVM was trained using the training collection to recognize Specific images. All the features values from the Specific images in the training collection were entered as positive examples to the SVM while all the feature values from the General images were entered as negative examples to the SVM. The SVM was then trained using different kernel functions such as linear and polynomial. The kernel function that performed best for this task was the radial basis function. The different parameters to use with this kernel were then optimized such as the cost factor for error and the gamma parameter for the kernel.

The SVM was trained and tested using all 6 image features. Once the SVM was trained, it was applied to the test collection. The SVM returns a confidence value that a certain image belongs to the Specific class. If this confidence value is greater than a threshold the image is classed as belonging to the Specific class of images. If the confidence value is below the threshold the image is classed as belonging to the General class of images.

The SVM that we trained classified 82.5% of the test collection correctly: 33 of the Specific images and 33 of General images in the test set were classified correctly.

3.3.2 K-Nearest Neighbour

We also implemented and trained a K-Nearest Neighbour classifier [7]. The K-Nearest Neighbour classifier was trained using the same 200 training images. Each test image was then run through the classifier and tagged as either Specific or General by the classifier.

We decided to use the K-NN classifier to test a number of different combinations of features to ascertain which combination of features would have the best classification performance and to discover which features helped discriminate well between Specific and General. These results are reported in Table 2.

Table 2. *K-Nearest Neighbour Classifier results broken down into Specific/General using different combinations of image features.*

Image Features Used	*Specific images classified correctly*	*General images classified correctly*	*Total images classified correctly*
All Features	80%	82.5%	81.25%
Without Facial Features	72.5%	70%	71.25%
Without Sharpness Variance	70%	100%	85%
Using Just 3 Modality Features	27%	95%	61%
Without 3 Modality Features	92.5%	47.5%	70%

From the reported results for the SVM and the results reported in Table 2, it is clear that it is possible to train computational classifiers to automatically recognize these image-text relations with reasonable accuracy based solely on low-level image features. Even though the highest overall result was obtained by using the K-NN classifier without the sharpness variance feature, the performance for classifying Specific images under this configuration was quite poor (70%). A more balanced result, which shows good performance for recognising both General and Specific, is more desirable therefore it seems that the support vector machine outperformed the K-NN marginally for this task.

4. Discussion

This work represents the first attempt to address image-text relations explicitly in both empirical and computational terms. We have found evidence that humans can predict something about the meaning of the text in a multimodal document by seeing only an accompanying image, and we have demonstrated that this prediction can be automated with a reasonable degree of success using only low-level image features. We are currently looking at how low-level text features can be used to make the reverse prediction. Based on preliminary research, it seems that when the Subject of the first sentence in a news story is a named person, then the accompanying photograph depicts that person's face large and full-frontal. We are also interested in whether other kinds of image-text relations can be classified automatically, such as those postulated and discussed in [1] and [2]. We expect that the recognition of image-text relations relies on a degree of conventionality in media production, so they will be more readily seen in mature forms, such as news, that are produced by trained professionals.

Knowledge of image-text relations could be applied to news adaptation and personalisation in a number of ways. Systems for indexing images on web pages rely on selecting keywords from the HTML text surrounding images [8]. The automatic classification of image-text relations should mean more reliable selection of keywords, e.g. in our cases when the classification is Specific then the first name in the news story should be used as an index term for the image, but not when the classification is General. When adapting and generating multimedia content automatically, better images to illustrate texts could be selected by consideration of image-text relations, e.g. to ensure that an image illustrating a text about a specific person shows their face large, centered and in sharp focus compared to the background.

More generally, in recent years there has been great interest in multimodal data fusion and multimedia information integration both for semantic media analysis and to assist in the creation and adaptation of multimedia content. In [9] the need to integrate textual information associated with images was recognised as a key strategy in closing the semantic gap. Text and image features have been fused for auto-annotation and auto-illustration [10, 11], for web image retrieval [12] and for web page retrieval [13], but none of this work has addressed the great variety of image-text relations that exist in real-world multimodal documents. The same can be said for attempts to index video data with associated text. Work on multimedia adaptation [14, 15, 16] has concentrated on the analysis of page layout but has not addressed the semantic nature of the relationships between different media items. We envisage all such work being enhanced by an appreciation for image-text relations in multimodal documents.

5. References

[1] R. Martinec, and A. Salway, “A System for Image-Text Relations in New (and Old) Media”, *Visual Communication* 4(3), 2005, pp. 337-371.

[2] A. Salway, and R. Martinec, “Some Ideas for Modelling Image-Text Combinations”, Dept. of Computing Technical Report CS-05-02, University of Surrey, 2005.

[3] S. Cooray, and N. O'Connor, “A Hybrid technique for face detection in colour images”, *AVSS - International Conference on Advanced Video and Signal based Surveillance*, Como, Italy, 15-16 September 2005.

[4] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, “A No-Reference Perceptual blur metric”, *ICIP 2002*.

[5] N. Cristianini, and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[6] SVMlight implementation, http://svmlight.joachims.org

[7] L. G. Sharpiro and G. C. Stockman, *Computer Vision*, Prentice Hall, 2001, pp. 103-104.

[8] J. Smith and S.-F. Chang, “Visually Searching the Web for Content”, *IEEE Multimedia* (4), 1997.

[9] A. Smeulders et al, “Content-Based Image Retrieval at the End of the Early Years”, *IEEE Trans. PAMI* 22(12), 2000.

[10] K. Barnard et al, “Matching Words and Pictures”, *Journal of Machine Learning Research* (3), 2003, pp. 1107-1135.

[11] K. Yanai, “Generic image classification using visual knowledge on the web”, *Procs. ACM Multimedia 2003*.

[12] T. Coelho et al, “Image Retrieval Using Multiple Evidence Ranking”, *IEEE Trans. KDE* 16(4), 2004.

[13] R. Zhao and W. Grosky, “Narrowing the Semantic Gap – Improved Text-Based Web Document Retrieval Using Visual Features”, *IEEE Trans. Multimedia* 4(2), 2002.

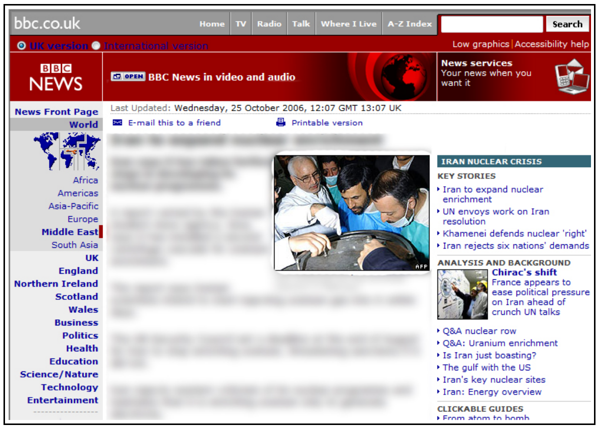
[14] Y. Chen et al, “Adapting Web Pages for Small-Screen Devices”, *IEEE Internet Computing* 9(1), 2005.

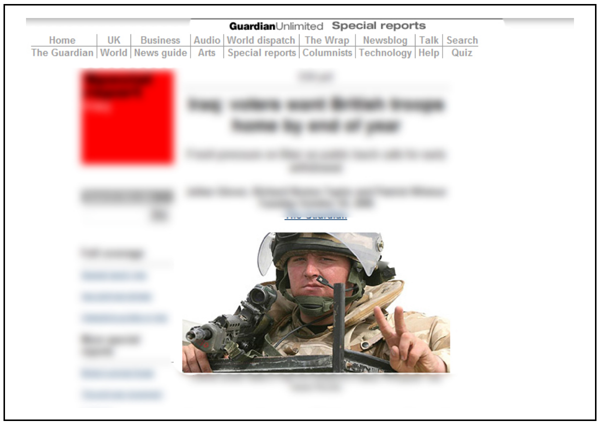
[15] D. Cai et al, “Extracting Content Structure for Web Pages based on Visual Representation”, *Procs. APWeb 2003*.

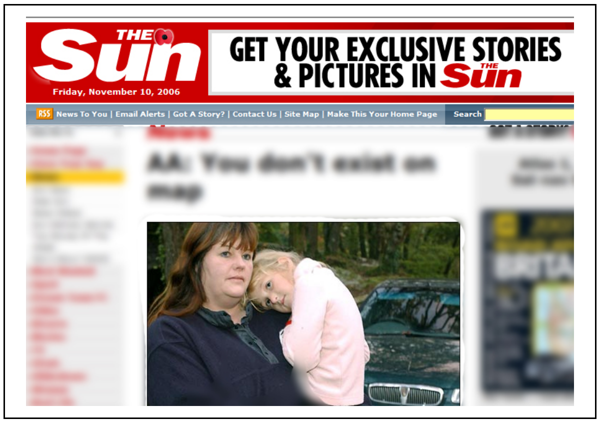
[16] R. Song et al, “Learning Block Importance Models for Web Pages”, *Procs. WWW 2004*.

Appendix

The three examples, discussed in Section 2, that were classified incorrectly by more than 4/25 subjects






THE Sun
Friday, November 10, 2006
GET YOUR EXCLUSIVE STORIES & PICTURES IN THE Sun
RSS
News To You | Email Alerts | Got A Story? | Contact Us | Site Map | Make This Your Home Page
Search