Abstract
Content-based indexing and retrieval of images and video requires a proper semantic description for image content. This paper discusses the mapping of high-level, application-specific, features to the visual primitives that are accessible through image processing techniques. A major difficulty is that there are currently no established methodologies for describing the contents of an image in terms of semantic features, and we suggest that semiotic approaches could be adapted to the task of image description in limited areas of expertise. After reviewing current trends in the mapping of high-level to low-level features, we present preliminary results that suggest a possible strategy for mapping semiotic descriptions to image processing primitives.

1 Introduction
Content-based indexing and retrieval of images and video is an essential condition for the development of a whole new class of interactive multimedia systems. This problem is now specifically addressed within the MPEG-7 initiative. In this sense, MPEG-7 differs significantly from the previous MPEG standards, as some of the content representation problems addressed by MPEG-7 belong to the field of Artificial Intelligence (AI). The description of high-level content features in MPEG-7 is faced with two major technical challenges. The first one is the proper description of content features for images and video, based on domain knowledge, for which there are currently no established methodologies. The second is the mapping of low-level image processing features to these high-level content descriptions, which is required to perform automatic extraction of high-level information from images.

In this paper, we will discuss some fundamental issues related to the mapping of image processing features to high-level symbolic features, which should also be related to image or video semantic descriptions. We will address these problems essentially from a content description standpoint, but will also support the discussion by presenting ongoing experiments in the automatic extraction of symbolic information from pictures in a restricted knowledge domain. Our test case deals with the automatic determination of the style categories for menswear images, where these categories are defined as a set of semantic features. We have chosen fashion pictures as a test case, because potentially these convey a large amount of both explicit and implicit information, ranging from direct visual properties to complex cultural connotations. Such an application is indeed part of the MPEG-7 list of potential applications [12], and is a genuine test case for the semantic representation of multimedia information. Through this toy example we will illustrate both semantic content description for images (section 2), and the possible automatic extraction of such information through image processing techniques (section 4).

2 Semantic Description of Images
Most of the current work in content-based image retrieval is still based on the direct use of visual primitives, such as colour, shape, texture and spatial features. For instance, the QBIC system [13] allows queries of large image databases based on visual image content, i.e. properties such as colour percentages, color layout, and textures occurring in the images. Such queries use the visual properties of images, so colours and textures may be
matched, and their positions, without describing them in words. VisualSEEK [16] is an automated scene analysis system, which uses visual properties of objects and their spatial layout, and employs colour sets as descriptive parameters. Mehtre et al. [9] recently described a composite colour-shape approach to image retrieval, based on shape and colour feature vectors. While their approach makes use of combinations of several features, it does not attempt at mapping them into high-level symbolic concepts, and queries are still based on “query images”. Descriptors associated with visual primitives are usually termed “low-level”, and, though higher-level descriptors could be obtained by combination of these basic descriptors, there is still a need to bridge the gap between these and the symbolic description of image content. Only the latter would eventually make it possible to perform queries on image databases using natural language, rather than sample images or image sketches.

A major problem for high-level image content description is the availability of a domain-related semantic description, as for the one produced by the semiotic analysis of images. In the following section, we will discuss possible directions within MPEG-7, and suggest that image content can be described on an a priori basis, very much in the way semantic content for words is described in AI, through semantic features. MPEG-7 documents [12] make explicit reference to both work on ontologies, like CYC [7] and natural language semantics, like WordNet [10]. Ontologies have a long history in philosophy, but in Artificial Intelligence they have come to acquire a more technical meaning; according to a definition by Tom Gruber, “An ontology is a specification of a conceptualisation”. Ontologies in AI have been developed for the purpose of knowledge interchange or the description of commonsense reasoning [7]. Yet, it is not clear how this work could apply to the description of images as cultural objects rather than descriptions of the real world. In many relevant applications (e.g., multimedia catalogues), image content is to be addressed in terms of the cultural information it conveys, as user preferences tend to be based on such cultural concepts. The formal description of content for cultural objects is mainly addressed by Semiotics, which for our purpose could be defined as a generalisation of semantics to other modalities than natural language, including images. One particular approach is called isologous semiotics, and assumes that image content descriptions can be based on the same kind of semantic features that describe the semantic content of words in sentences and texts. Whether or not this is true as a general rule, it can serve as good methodological approximation, and would certainly help in priming content description of images. In the case of fashion objects, such a hypothesis has been extensively developed by Barthes [3], in his classical work on the systematic description of the semiotic codes of fashion. Semantic content description is best performed through the use of semantic features, which description is compatible with highly contextual information [4]. There is a long history for features in the formal description of cultural material, mainly coming from structural linguistics [15]. Nevertheless, the isologous semiotics hypothesis does not imply that the high-level image description features should be derived from the analysis of natural language descriptions of these images. It rather indicates a commonality in terms of description methodology, together with the fact that there could be some overlap between the image features and some linguistic features, which would make their integration in practical applications easier. Whenever semiotics is to borrow its methodology from semantics, it would be based on a small set of descriptive operations, like the collection of meaning classes, and the opposition between elements of meaning within such classes. That is to say, for the sake of description, images sharing a possible common high-level category should be grouped together, and their content features would then be described by contrasting some images with others. There is, anyway, considerable material on the description of clothing elements, its significance, its meaning and connotation, such as in [3], or on a more humoristic perspective, but not without insight, in [14]. For instance, the formal shirt colour is /white/ by contrast with other colours. Classic ties usually follow striped or small regular patterns, and are restricted to a limited number of colours (See for instance: “Conservative ties with subtle patterns” from http://www.as.cmu.edu/career/search/dress.html). The aperture of a jacket generally varies according to the dress code; a deep aperture would indicate less formal dressing. Other concepts have been described, such as lapel size,

![Fig. 1. Descriptors used for dress code analysis](attachment:image.png)
which has been shown to be correlated with tie shape [14].

These descriptions are, however, valid within a determinate context only. This context accounts for the fact that features do not appear to combine freely. In classic dress, everything is classic, and in casual dress, formal elements are an exception.

We can now illustrate this approach by way of a description of some menswear photographs; we have selected several descriptors related to the jacket cut (like single/double-breasted or the number of buttons), or the colour and patterns of various elements (shirt, tie, jacket), and mapped these into three main categories for a dress code, which are: formal, classic, and modern (see fig. 1).

The high-level domain concepts used in the content description of fashion images can be characterised through the following properties:

- **Contextual interpretation.** A single feature cannot be interpreted out of a more global context, which is determined by the occurrence of other features. For instance, considering our “aperture” dimension, a feature like /deep aperture/ cannot be always be considered as an index for /classic/, since it could appear on a smoking jacket. In that case, the contextual feature will override the specific one.

- **Time validity of feature descriptions.** Classic style is, by definition, subject to little variations within reasonable time spans. Conversely, a /modern/ category description has a limited validity over time.

- **Prototype effects.** Just as with other forms of cultural descriptions, some generic notions may display prototype effects related to feature typicality, or the default values for a given notion.

- **Discriminating power.** Generally speaking, the discrimination power of an isolated feature tends to be low. For instance, it is not possible to use /dark suit/ as a discriminative feature, as it can be part of both the /formal/ and the /modern/ categories. Potentially, this could differ from linguistic semantic descriptions.

Taking as a starting point the high-level content description for these images, we will now discuss how these could be mapped to image processing features produced by various segmentation and analysis techniques. These features include: contours, areas, textures, colour content, brightness, relative position (of shapes), sizes (e.g., of lapels, aperture), and relative distribution of colours within specific regions, etc. However, in order to give directions for such a mapping, we need to review firstly some current conceptions concerning the relationship between high-level and low-level features.

3 Mapping Visual Features to Semantic Features

In this section, we review current trends in the mapping of low-level visual features to high-level domain features. We will classify these as direct mapping, template-based mapping, and hierarchical mapping.

**Direct mapping** from image processing features to high-level features consists in the use of a visual feature to infer directly a high-level domain concept, or to discriminate between two high-level categories. A first example is the existence of “characteristic” content features that have a direct mapping in terms of image processing properties. By characteristic, we mean that this high-level feature does not require a complex contextual interpretation, and that its stand-alone recognition can serve as the sole basis for classification or inference. One application, in the case of dress code identification, would be the recognition of a “jeans” texture, which would directly classify the trend into a casual category. Another instance of direct mapping is the possibility of discriminating between categories on the basis of a single salient property of an image or a video sequence. For instance, Vasconcelos and Lippman [18] have been able to use simple video image properties, like local activity of a sequence and its duration, to discriminate between /action/ and /non-action/ movie categories.

**Template-based mapping** relies on templates that can be directly applied to the image. For instance, one could apply a filter on a determinate image region and further count for specific patterns detected within that region, or determine their relative positions. In our fashion application, this could be, for instance, applied to the detection of buttons on a jacket. From a more technical perspective, definition of objects and shapes in images can be accomplished by means of breaking them down into primitives, and then assigning a code to describe the interconnection of these primitives. Edge information may be used to define the boundaries. Edge detection filters are often efficient for detecting segments, but more sophistication is necessary to
close contours and produce well-defined shapes, which can then be identified in terms of the domain knowledge. However, there is also evidence that extensive segmentation (including closing of contours) is not an absolute requirement for further processing, such as the recognition of selected patterns by alignment [17]. Other image attributes may be determined, for example, the relative quantities of primary components (R,G,B, or C,M,Y,K) in the image, and their locality (especially within a given region delimited by a template). Simple versions could be devised in the case of images depicting human subjects, by labelling regions of interest according to a domain-dependent description scheme, this labelling making in turn further image segmentation more accessible. For instance, many fashion pictures featuring models could be simply tagged according to the various body segments.

This would enable a mapping between the fashion concepts and geometrical computations over particular regions based on image processing functionalities, to be interpreted in terms of content features. Such a segmentation would be generic, and could be applied to any picture representing a human subject (including other applications such as in sports). Once the head location becomes available, a feature like /aperture_depth/ would be easy to compute by processing the selected image region (see e.g. fig. 1 and 4). Such an approach is valid only insofar as a class of pictures displaying human subjects is of general interest for many applications, but this is very likely to be the case in terms of the application to fashion/clothing, sports, movies, advertising, health & fitness, etc. Indeed, human activity is a major category of image and video databases [8]. Another related approach is model-based vision, which relies on a tridimensional object model to perform image processing and interpret further the results obtained. This technique has been successfully applied to the detection of moving vehicles in images [6]. However, in systems implemented so far mapping tends to take place between models’ visual features and similar image features, without explicit representation of high-level features.

Finally, hierarchical mapping assembles features into different layers. In each of these layers, features are described in terms of lower-level features, while still retaining a proper interpretation within that layer. An example of hierarchical mapping is the use of state transition models to recognise objects in sport scenes [11]. For instance, when analysing a soccer game image, a /player/ instance can be recognised from the occurrence of /face/ and /hair/ features, which consitute an intermediate description layer. This layer can be mapped directly to image processing features, for instance those computing skin-colour measurements [8] that account for the recognition of the /face/ descriptor. Such a system can work on both positive and negative evidence when it comes to differentiating a player from his background. Hierarchical mapping appears to require an associated set of procedures to implement the strategy by which low-level features are used to produce high-level descriptors. It can also be noted that some layers of description in hierarchical mapping can make recourse to templates for feature determination.

In the next section, we will describe in greater detail a possible approach to hierarchical mapping for the automatic determination of dress code from images, based on our previous content description, as well as image segmentation results. We will suggest that most of the symbolic features described above can actually be obtained through relatively simple image processing techniques. As a consequence, the design of an image retrieval system will strongly rely on the mapping procedure.

### 4 The hierarchical approach to feature mapping

We can now address the problem of mapping semantic features to image processing composite features in our sample images. The task we want to illustrate is the automatic determination of the fashion style and dress code for menswear (suits, in this particular case). We will rely on our sample classification presented in table 1.

<table>
<thead>
<tr>
<th></th>
<th>Formal</th>
<th>Classic</th>
<th>Modern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jacket</td>
<td>Double breasted</td>
<td>Two buttons</td>
<td>Three buttons</td>
</tr>
<tr>
<td>Shirt colour</td>
<td>White</td>
<td>White, light blue</td>
<td>Dark / Black</td>
</tr>
<tr>
<td>Jacket colour</td>
<td>Black</td>
<td>Grey, Navy …</td>
<td>Dark</td>
</tr>
<tr>
<td>Aperture</td>
<td>Medium</td>
<td>Deep</td>
<td>Small</td>
</tr>
<tr>
<td>Tie</td>
<td>Small patterns</td>
<td>Regular patterns</td>
<td>Larger patterns</td>
</tr>
</tbody>
</table>

Table 1. Semantic description of a sample dress code.
The top level of description is constituted by the category to which the suit depicted belongs, such as: /formal/, /classic/, /modern/. This category classifies the image as a whole and can be used to index the image or to retrieve it in response to a query. It should be noted that, in the case of natural languages queries, these can in theory be of arbitrary complexity and require inference, for instance, determining the appropriate style as a function of explicit characteristics of the user, such as: age, working environment, income, etc.

Intermediate level domain concepts describe image properties, which are explicitly referred to for the classification of a given suit. As such, they are still domain concepts, and their description as image properties belongs both to application knowledge and image semiotics. These can be general, such as /shirt colour/, or more specific, such as /aperture depth/, /jacket-tie coordination/, /lapel size/, etc. A sample set can be generated from table 1, and would include: /jacket cut/, /jacket colour/, /shirt colour/, /tie patterns/, /tie colour/, /aperture/, etc.

These concepts would map to a high-level image processing layer, dedicated to complex operations such as measuring the size of a given area, computing the colour density of the same area, counting items in an image region, etc. For instance, the colour of the shirt could be determined through the color density of the external region of the aperture, and the jacket cut could be determined by counting the number of buttons. This stage is an essential step in bridging the gap between high-level features and image processing features. It can be described as a procedural mapping, which associates a set of image processing operations to a given high-level descriptor. For instance, determining the colour of the shirt in a fashion image can be decomposed into several steps: i) detecting the aperture and tie regions through edge detection ii) performing colour density measurements on the area delimited by these regions. Though details may vary in each picture, the procedure would be essentially the same.

These computations require elementary features acquired by basic image processing techniques such as segmentation and/or image density measurements. We have used a novel edge detection technique, which has been developed by one of the authors (RG) and perform analysis of images using hexagonal pixel formats [5], in which the Cartesian coordinate system is replaced by the tri-axis approach, with 60 degrees between the axes. The advantages of this includes better resolution along objects with curved boundaries, improved processing speed, and better following of the direction of curved boundaries. In addition, novel filters for noise removal can also be generated [1] [2]. The hexagonal format of pixel used for image segmentation also leads to lower errors and better joining of line segments than the standard approaches to boundary extraction.

Figures 3 shows the result of edge detection for the three suits in figure 2 (a-b-c). These results suggest that most of the relevant concepts described above are accessible to simple edge detection techniques. The jacket aperture, and, subsequently, the tie, would be easily delimited, though this is more difficult in the case of low contrast between jacket and tie. However, even in this case, it should still be possible to generate useful information. We have seen that counting the number of buttons could be an easy way to determine the jacket cut: figures 4 (a-b-c) demonstrate, regardless of the jacket colour, that buttons can be specifically tracked in predetermined regions. However, the lapel (and its size) was not successfully detected in any of the images, due to unsufficient contrast and perhaps illumination conditions. This might be better determined from moving image sequences. From the segmentation information obtained, it would still be possible to determine the respective regions for aperture and tie, which would give access to the following domain features : /jacket cut/, /jacket colour/, /shirt colour/, /shirt-tie coordination/, /aperture/, etc. by applying appropriate colour density measurements within these regions.

By performing the above steps, we would obtain the following classifications, which do correspond to the original descriptions:

- For figure 2a: /double_breasted/ + /jacket = black/ + /shirt = white/ + /medium aperture/, a description compatible with the /formal/ category
- For figure 2b: /single_breasted/, /two buttons/, /deep aperture/, /shirt = light blue/, which is compatible with the /classic/ dress code
- For figure 2c: /single_breasted/, /three buttons/, /jacket = navy/, /shirt = dark/, indicating a /modern/ category for the suit.

2 At this stage, there would be no detection of shirt patterns.
However simple, this example would still demonstrate that a mapping between semiotic features for images and visual processing can be properly specified within a given application, hence making semiotic description of images a good starting point for image content description, especially for those applications requiring highly contextual knowledge. The hierarchical mapping ensures a proper balance between top-down and bottom-up processes in image analysis, as pure bottom-up approaches tend to suffer from various limitations [17].

5 Conclusions and future work

We have described a possible approach to the mapping of image processing features onto high-level content features. This approach is based on the recognition of the need for specific domain knowledge features, which come as a semiotic description of the image content. This description should be enhanced by a certain number of inference rules for the interpretation of isolated content features, which would be mapped onto low-level features, into significant domain concepts. These high-level concepts would be the ultimate product of indexing and retrieval, and are likely to be interoperable with other representations, such as natural language queries. The optimal hierarchical mapping process is still to be determined. It could be implemented as a knowledge-based system, taking as input a manual description of the semantic features involved in fashion and dress codes, in the spirit of Barthes [3]. Such a knowledge-based system would allow exploratory programming of several mapping strategies, and appears more appropriate than statistical approaches, as the available knowledge is essentially qualitative in nature.

Acknowledgements

Syeed Abu Bakar, Rami Qahwaji and Michael McNeill are thanked for technical assistance. We also wish to thank Carlton Reeve and David Cole.

References


Figure 2 (a-b-c). Suit images used in the experiment

Figure 3 (a-b-c). Edge detection through Hexagonal Pixel Format technique

Figure 4 (a-b-c). Descriptive features are concentrated on specific image regions