UNIVERSITY OF AMSTERDAM

Faculty of Science

Exam

Media Understanding
Bachelor Kunstmatige Intelligentie

Resit Exam

Date: May 30, 2017
Time: 13:00-16:00
Place: A1.04

Number of pages: 8 (including front page)
Number of questions: 6

BEFORE YOU START

- Please **wait** until you are instructed to open the booklet.
- Check if your version of the exam is complete.
- Write down **your name, student ID number,** and if applicable the **version number** of the exam **on each sheet** that you hand in. Also **number the pages.**
- Your **mobile phone** has to be switched off and in the coat or bag. Your **coat and bag** must be under your table.
- **Tools allowed:** ruler, graphics calculator

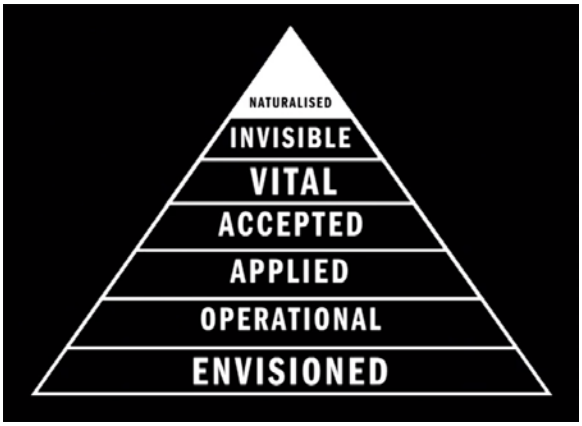
PRACTICAL MATTERS

- The first 30 minutes and the last 15 minutes you are not allowed to leave the room, not even to visit the toilet.
- You are obliged to identify yourself at the request of the examiner (or his representative) with a proof of your enrollment or a valid ID.
- During the examination it is not permitted to visit the toilet, unless the invigilator gives permission to do so.
- 15 minutes before the end, you will be warned that the time to hand in is approaching.
- If applicable, please fill out the evaluation form at the end of the exam.
- You may answer the questions both in **Dutch** or **English**.

Good luck!

References

[1] Horst Eidenberger, 'Fundamental Media Understanding', 2nd edition, Books on Demand GmbH, Norderstedt, Germany, 2011, ISBN: 978-3-842-37917-6.
[2] Wood, J., et al. "Riding the wave: How Europe can gain from the rising tide of scientific data." Final report of the High level Expert Group on Scientific Data–A submission to the European Commission, European Union." May 13 (2010).
[3] Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., & Milford, M. J. (2016). Visual place recognition: A survey. IEEE Transactions on Robotics, 32(1), 1-19.
[4] H. Bay, T. Tuytelaars, and L. Van Gool, " SURF: Speeded up robust features," in Proc. Eur. Conf. Comput. Vis., 2006, pp. 404–417.
[5] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global mage features in recognition," in Visual Perception – Fundamentals of Awareness: Multi-Sensory Integration and High-Order Perception. New York, NY, USA: Elsevier, 2006, pp. 23–36.
[6] Rocamora, M., Cancela, P., & Pardo, A. (2014). Query by humming: Automatically building the database from music recordings. *Pattern Recognition Letters*, *36*, 272-280.



Pyramid of Technology, Koert van Mensvoort, 2014.

Question 1

The EU estimates that 30% of the world storage is used for medical imaging. Those images could be acquired by nuclear imaging, X-rays or magnetic resonance. The format of the images could be simple 2D (e.g. dermatography) to complex 3D + time (e.g. functional MRI). Yet, also for medical imaging this multimodal information is only valuable once the images are interpreted. To make sense of the medical images on tries to describe them with higher level concepts, such as the Bag of Visual Words approach.

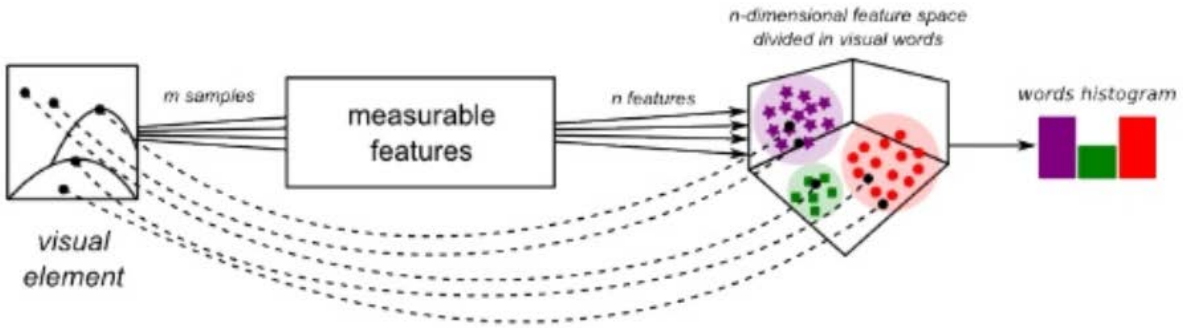


Fig. 1a: The Bag of Visual Words approach (Courtesy Antonio Foncubierta Rodriguez, Medical Image Analysis and Retrieval at ETH Zurich)

In the book of Horst Eidenberger [1] the Bag of Visual Words is generalized to a Bag of Features method.

a) Could you describe the Bag of Features approach in your own words?

The Bag of Features learns the occurrence of certain features for a set of visual words. The presence of features of a query element is then matched against the visual words.

The Bag of Features method is somewhere in the middle of a dynamic association method and a traditional distance-based method.

b) Could you describe the difference in optimization on element level and set level for the Bag of Features method?

The optimization on element level is to minimize the distance or maximize the similarity in the multidimensional feature space. The optimization on set level is to learn the number of sets (clusters) and how to distinguish them.

In the feature space of Fig 1a three clusters of visual words are depicted. The categorization is clearly performed by hedging.

c) Could you indicate from the following categorization methods if they are hedgers or separators?

- Cluster analysis Hedger
- VSM classifier Hedger
- K-means classifier Seperator
- K-nearest neighbor classifier Hedger

See page 198.

Yet, segmentation of the image remains important in medical image interpretation. MRI images do not have color, so only texture and shape are available as visual cues. Antonio Foncubierta Rodriguez performs segmentation based on texture (see Fig 1b).

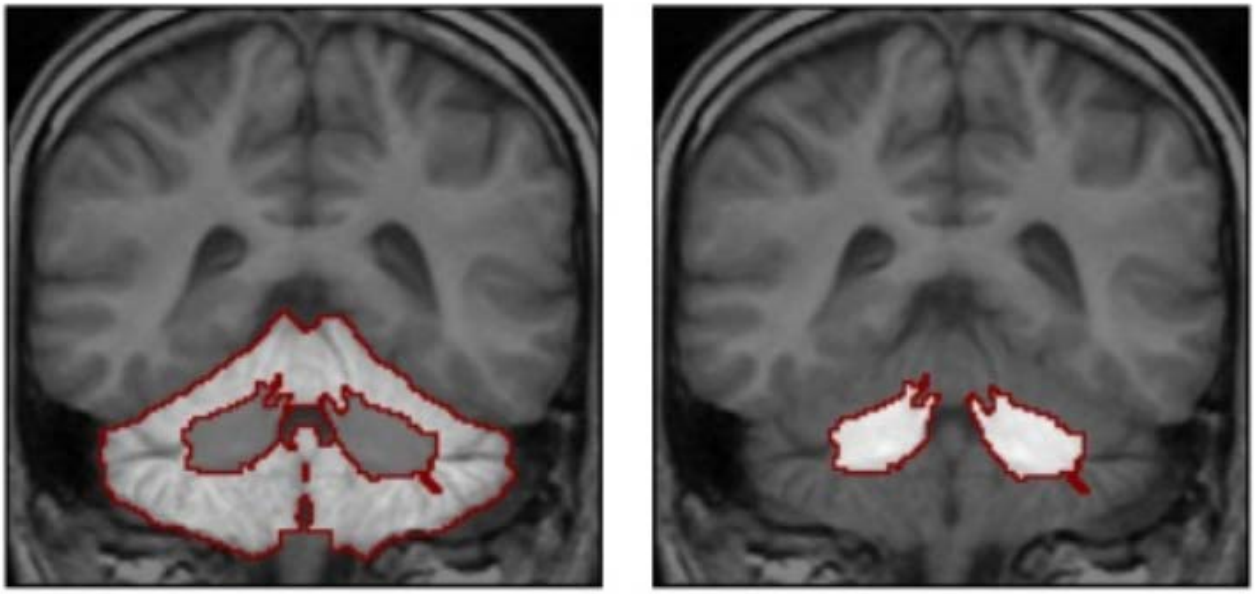


Fig. 1b: MRI scan of the brain with segmentations of the Cerebellum cortex and white matter. (Courtesy Antonio Foncubierta Rodriguez, Medical Image Analysis and Retrieval at ETH Zurich)

Courseness, regularity and directionality are three very important properties of textures.

d) Describe how coarseness, regularity and directionality could be measured from a grey-scaled image?

See page 88.
coarseness: comparing the statistical moments at different resolutions
regularity: comparing coarseness in neighboring cells
directionality: comparing coarseness in neighboring cells only in certain directions

Question 2

A traditional distance-based method for text retrieval is the vector space model , for instance based on the cosine similarity.

a) What is the benefit of using the cosine similarity compared with using the Euclidian distance?
Explain your answer!

See slide 19 of Bob van der Velde.
Cosine similarity is inherently normalized & outliers have less impact.

b) Could one use such the cosine similarity also in a global context; for a bigger collection of documents (corpus)?
Explain your answer!

See slide 22 of Bob van der Velde.
Similarity scores can only be used relatively, for thresholds you should group them in topics

c) Can you explain how topics can be used to improve the interpretation of text?

See slide 23 of Bob van der Velde.
Topics could be used to learn the context and the associated frequencies of words in that context.

Question 3

Visual place recognition is a well-defined but extremely challenging problem to solve [3].

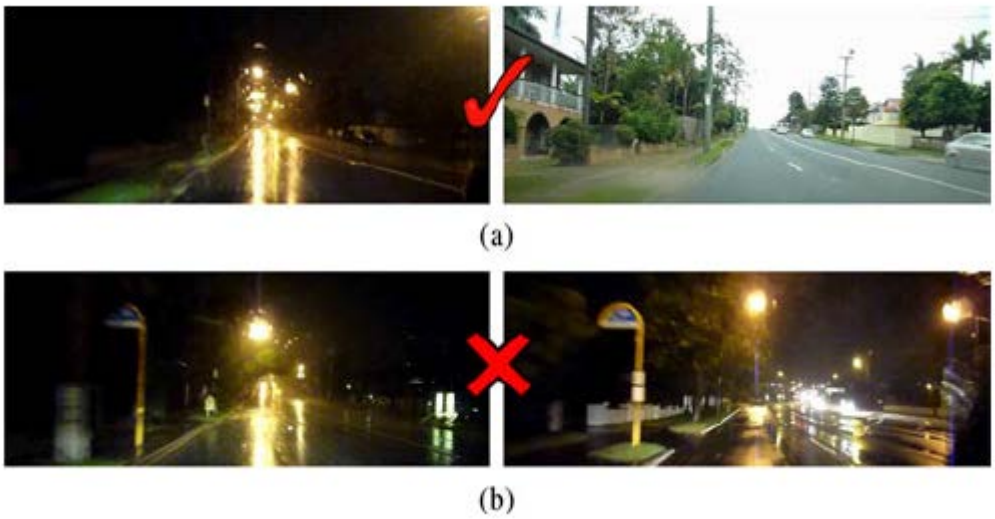


Fig 3a: Visual place recognition must be able to (a) successfully match very perceptually different images while (b) also rejecting incorrect matches between aliased image pairs of different places. (Courtesy Stephane Lowry et al. [3].)

Visual place description techniques fall into two broad categories: those that selectively extract parts of the image that are in some way interesting or notable; and those that describe the whole scene, without a selection phase (See Fig 3b).

Fig 3b: Visual place description techniques fall into two broad categories. (a) interesting or salient parts of the image selected for extraction. (b) described the whole scene in a predefined way such as the grid shown (Courtesy Stephane Lowry et al. [3].)

An examples of the first category are SURF features [4]; an example of the second category are whole image descriptors as GIST [5].

In Chapter 11 of the book of Horst Eidenberger [1] the building blocks of feature extraction are indicated (see Fig 1a). For every type of media this feature extraction is slightly different.

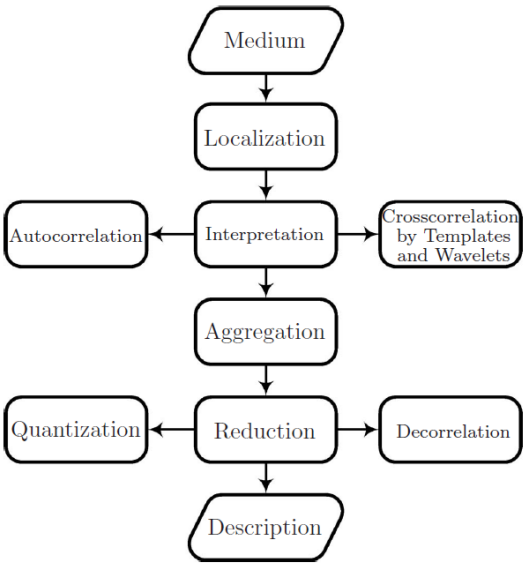


Fig 3c: The feature extraction building blocks (Courtesy Horst Eidenberger [1]).

As can be seen in the scatter plot of Fig 3c there are different regions which correspond to pixels with different semantic meaning in the image. Pixels in the green region correspond to vegetation and pixels in the blue region correspond to sky.

a) Could you indicate in which building block that generates the *interesting or salient parts of the image selected for extraction*?
Explain your answer!

See section 11.2 of the book of Horst Eidenberger [1].
Finding interesting points is autocorrelation (finding edges, corners, …). Yet only the distinctive points are kept, which includes considering the similarities between descriptions and removing the redundant parts (decorrelation filter).

b) Could you indicate in which building block that generates the *described the whole scene in a predefined way such as the grid shown* are generated?
Explain your answer!

See section 11.2 of the book of Horst Eidenberger [1].
Generating a grid is localization (decomposition into media chunks). The description of this media chunks with features is again interpretation (with autocorrelation).

When an image contain too many local features, direct matching of all those local features can be inefficient. The bag-of-word model increase efficiency by quantizing local features in a vocabulary. A typical vocabulary of visual words contains 5.000 – 10.000 words, but also vocabularies as large as 100.000 words have been used. Effectively in a visual words are reduced to binaries strings or histograms of length *n,* where *n* is the number of words in the vocabulary.

Images described using the bag-of-words model can be efficiently compared using binary string comparison such as the Hamming distance. The Levenshtein distance is a generalization of the Hamming distance

c) Can you explain binary string comparison like the Levenshtein distance?

See page 151 of the book of Horst Eidenberger [1].
The Levenshtein distance measures the difference between two strings by counting how many insert, delete and substitute operations are needed to transform one string in another string.

Also global place descriptors use a wide variety of image features – such as edges, corners and color patches – combined in a fingerprint of a location. By ordering these features in a sequence between 0^{o} and 360^{o}, also global place recognition can be reduced to string matching. The popular whole-image descriptor GIST [5] uses Gabor filters at different orientations and different frequencies to extract information from the image.

d) Gabor wavelets are also described in Chapter 5 of the book of Horst Eidenberger [1]. Which general property of an image can be categorized by a Gabor wavelet?

See page 90 of the book of Horst Eidenberger [1].
Gabor wavelets can be used to describe texture.

Question 4

To measure melody similarity, Rocamora *et al* [6] extend the Levenshtein distance with duration and pitch information. Duration and pitch are two of the perceptual properties of sound.

a) What are the other three fundamental properties of sound?

See page 59 of the book of Horst Eidenberger [1].
Loudness, Rhythm & Timbre

There is a difference between the perceived pitch by a human and the measured frequency of the sound.

b) Describe in your own words the relation between the perceived pitch and the measured frequency.

See page 61-62 of the book of Horst Eidenberger [1].
Log-log relationship which is non-linear. Below 1 kHz the perceived pitch increase over-linearly, above 1 kHz under-linearly.

c) What is the unit of perceived pitch?

See page 61-62 of the book of Horst Eidenberger [1].
Mel

Patterns of changes in the pitch can be measured by Mel-Frequency Cepstral Coefficients (MFCC).

d) Which fundamental properties of sound can be estimated with MFCC?

See slide 20 of Anders Bouwer
MFCC is used to estimate timbre, the variations in pitch

Question 5

Emotions can be measured by interpreting many different media and signals. Within the area of Affective Computing, three perspectives can be distinguished on how to compute with emotions.

a) Give these three perspectives, and explain each of them in one sentence.

See slide 10 of Tibor Bosse
Categorical perspective (assume a fixed set of basic emotion categories)

Dimensional perspective (view emotions as points in a continues space)
Componential perspective (emotions are produced by a combination of components)

Emotions can be classified in the dimensional perspective in a 3D space.

b) Which three numerical dimensions are often used to represent all possible emotions?

See slide 12 of Tibor Bosse:
valence, arousal, dominance (or pleasure, arousal, dominance)

c) Give a positive and negative emotion for each dimension.

See slide 13-14 of Tibor Bosse:
arousal: sleepy (negative) – surprise (positive)
valence: sad (negative), happy (positive),
dominance: fear(negative) – anger (positive)

In the OCC-Model of Ortony, Clore, and Collins one not only focusses on the consequences of events for one self, but also on the emotional reaction on other consequences, actions and aspects.

d) Which other consequences, actions and aspects? Who or what are involved?
Give an example of one consequence, action and aspect.

See slide 16 of Tibor Bosse:
consequences for others
actions of agents
aspects of objects

Question 6

One can simply express the meaning of a conditional probability textually in the following way:

$$P(y|x) = \frac{number\ of\ times, events\ x\ and\ y\ happen\ together}{number\ of\ times, event\ x\ happens}$$

a) How would you formulate textually the conditional probability *P(x|y)*?

See page 157 of the book of Horst Eidenberger [1].:
number of times,events x and y happen together / number of times,event y happens

By combining the relation between the joint probability *P(x,y)* with both the conditional probabilities *P(x|y)* and *P(y|x)* one can derive the famous Bayes theorem.

b) Please make this derivation.

See page 157 of the book of Horst Eidenberger [1].:
P(x|y) = P(x,y)/P(y) => P(x,y) = P(x|y) P(y)
P(y|x) = P(x,y)/P(x) => P(x,y) = P(y|x) P(x) // Chain rule

=> P(x|y) P(y) = P(y|x) P(x) => P(y|x) = P(x|y) P(y) / P(x)

You reached the end of the exam!