UNIVERSITY OF AMSTERDAM

Faculty of Science

Exam

Media Understanding
Bachelor Kunstmatige Intelligentie

Regular Exam

Date: March 31, 2017
Time: 9:00-12:00
Place: Universitair Sport Centrum, Sporthal 1

Number of pages: 10 (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

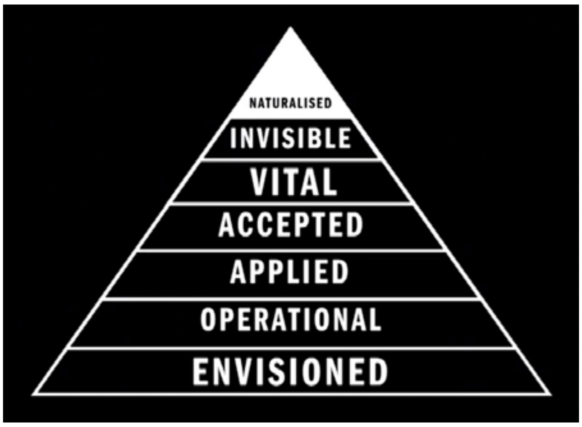
[1] Horst Eidenberger, 'Fundamental Media Understanding', 2nd edition, Books on Demand GmbH, Norderstedt, Germany, 2011, ISBN: 978-3-842-37917-6.

[2] Evgeniy Gabrilovich and Markovitch Shaul. "Computing semantic relatedness using Wikipedia-based explicit semantic analysis." IJcAI. Vol. 7. 2007.

[3] Abhinav Valada, Gabriel L. Oliveira, Thomas Brox, and Wolfram Burgard "Deep Multispectral Semantic Scene Understanding of Forested Environments using Multimodal Fusion". The International Symposium on Experimental Robotics (ISER 2016). October 2016.

[4] David M. Bradley, Scott M. Thayer, Anthony Stentz, and Peter Rander, "Vegetation Detection for Mobile Robot Navigation", Technical Report CMU-RI-TR-04-12, Carnegie Mellon University, February 2004.

[5] DeLiang Wang, "Deep learning reinvents the hearing aid," in IEEE Spectrum, vol. 54, no. 3, pp. 32-37, March 2017.



Pyramid of Technology, Koert van Mensvoort, 2014.

Question 1

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.

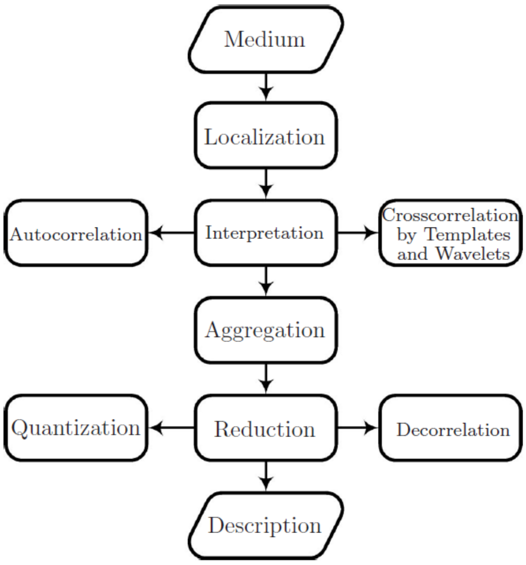


Fig. 1a: The feature extraction building blocks

One of those media are stock signals (see Fig 1b). In this figure some common used terms in stock chart analysis are indicated

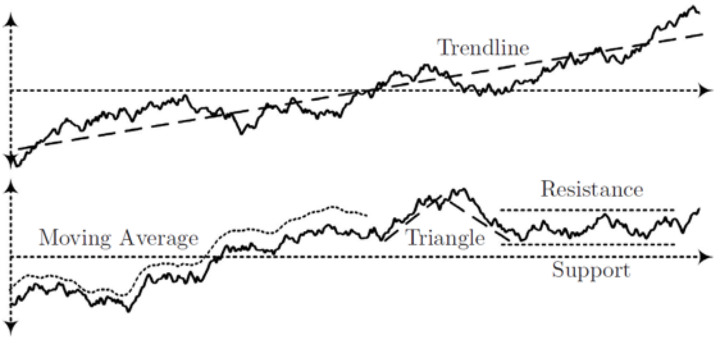


Fig. 1b: Feature of technical stock chart analysis

a) Could you indicate in which building block the *Trendline* is estimated? **Explain your answer!**

Recognizing the Trendline is taking the Statistical Moment of an already Localized timewindow, so it is part of the Interpretation block (see Table 11.1 on page 202). Aggregation is when multiple objects (stock charts / time windows) are combined.

b) Could you indicate in which building block the *Triangle* pattern is recognized? **Explain your answer**.

Recognizing the Triangle pattern is Template Matching (8), a form of cross-correlation (8) – both terms mentioned (10). Cross-correlation is part of interpretation, yet that claim is less specific (6)

Question 2

The *bag-of-words* method is a middle ground between traditional distance-based methods and dynamic association methods. The matching of two objects can be guided by similarity metrics such as the Euclidian distance or the cosine similarity.

a) Are the Euclidian distance and the cosine similarity positive or negative convolutions? **Explain your answer!**

A negative convolution becomes minimal for identical objects, as the Euclidian distance. The cosine similarity is based on the inner product and becomes 1 for identical objects (so a positive convolution).

See slide 19/35 from Text-slides. See also page 70 of the book of Horst Eidenberger [1].

The dynamic association is not on object level but is applied when groups of elements are formed (phrases) and optimization is performed on group level. For information retrieval, that means clustering of words into topics. One way of clustering words into topics is by creating a list of concepts from Wikipedia (see Fig 2a)

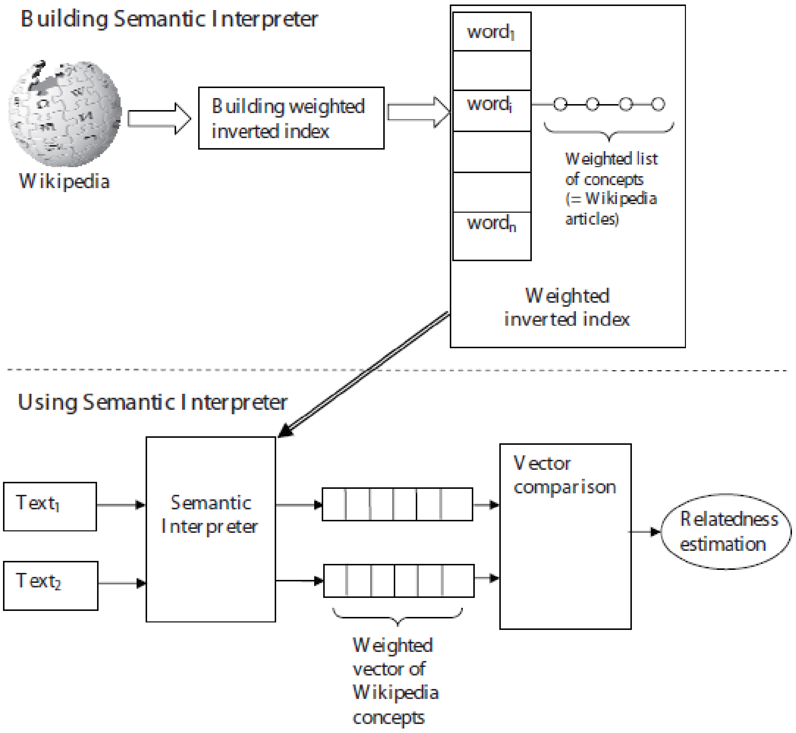

Building Semantic Interpreter
Wikipedia
Building weighted inverted index
word_1
word_i
word_n
Weighted list of concepts (= Wikipedia articles)
Weighted inverted index
Using Semantic Interpreter
Text_1
Text_2
Semantic Interpreter
Vector comparison
Relatedness estimation
Weighted vector of Wikipedia concepts

Fig. 2a: Building a semantic interpreter from Wikipedia concepts [2]

b) Could the result of the Semantic Interpreter, the relatedness estimation, be seen as the result of template matching on Wikipedia concepts?
Explain your answer!

No, The Wikipedia are not used as templates, but as reference points (compare the Triangle Equality on page 191 of the book of Horst Eidenberger [1]).

The text of the Wikipedia articles *t* could also be used to as reference to combine a *human choice model* with edit distances between two text-fragments *x* and *y*

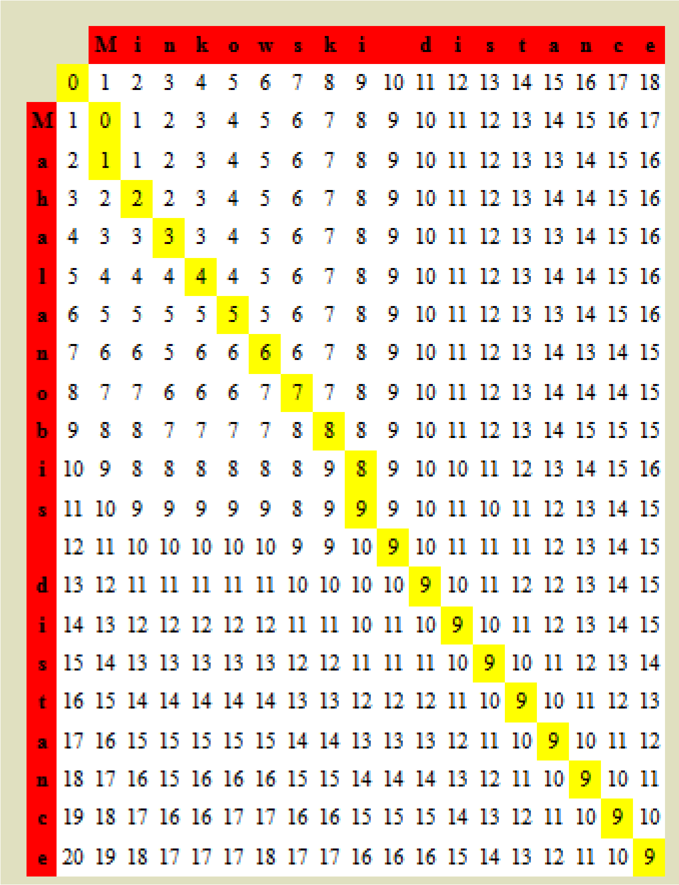
$$m_{edit}(x,y) = \frac{m_{edit}(t,y)}{m_{edit}(t,x) + m_{edit}(t,y)}$$

The edit distance can be measured with the Levenshtein metric with the operations *substitute, insert, delete.*

c) Please calculate the Levenshtein distance between the text fragments “Mahalanobis distance” and ”Minkowski distance”
Show the steps of your calculation!

The Levenshtein distance is 9. The Levenshtein distance is described at page 151.

Calculation can be checked with http://www.let.rug.nl/~kleiweg/lev/



		M	i	n	k	o	w	s	k	i		d	i	s	t	a	n	c	e
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
M	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
a	2	1	1	2	3	4	5	6	7	8	9	10	11	12	13	13	14	15	16
h	3	2	2	2	3	4	5	6	7	8	9	10	11	12	13	14	14	15	16
a	4	3	3	3	3	4	5	6	7	8	9	10	11	12	13	13	14	15	16
l	5	4	4	4	4	4	5	6	7	8	9	10	11	12	13	14	14	15	16
a	6	5	5	5	5	5	5	6	7	8	9	10	11	12	13	13	14	15	16
n	7	6	6	5	6	6	6	6	7	8	9	10	11	12	13	14	13	14	15
o	8	7	7	6	6	6	7	7	7	8	9	10	11	12	13	14	14	14	15
b	9	8	8	7	7	7	7	8	8	8	9	10	11	12	13	14	15	15	15
i	10	9	8	8	8	8	8	8	9	8	9	10	10	11	12	13	14	15	16
s	11	10	9	9	9	9	9	8	9	9	9	10	11	10	11	12	13	14	15
	12	11	10	10	10	10	10	9	9	10	9	10	11	11	11	12	13	14	15
d	13	12	11	11	11	11	11	10	10	10	10	9	10	11	12	12	13	14	15
i	14	13	12	12	12	12	12	11	11	10	11	10	9	10	11	12	13	14	15
s	15	14	13	13	13	13	13	12	12	11	11	11	10	9	10	11	12	13	14
t	16	15	14	14	14	14	14	13	13	12	12	12	11	10	9	10	11	12	13
a	17	16	15	15	15	15	15	14	14	13	13	13	12	11	10	9	10	11	12
n	18	17	16	15	16	16	16	15	15	14	14	14	13	12	11	10	9	10	11
c	19	18	17	16	16	17	17	16	16	15	15	15	14	13	12	11	10	9	10
e	20	19	18	17	17	17	18	17	17	16	16	16	15	14	13	12	11	10	9

M —a —h —a —l —a —n —o —b —i —s — —d —i —s —t —a —n —c —e +1 = 1 (substitude)
M — —i —n —k —o —w —s —k —i — — —d —i —s —t —a —n —c —e

M —a —h —a —l —a —n —o —b —i —s — —d —i —s —t —a —n —c —e +1 = 2 (substitude)
M —i — —n —k —o —w —s —k —i — — —d —i —s —t —a —n —c —e

M —a —h —a —l —a —n —o —b —i —s — —d —i —s —t —a —n —c —e +1 = 3 (substitude)
M —i —n — —k —o —w —s —k —i — — —d —i —s —t —a —n —c —e

M —a —h —a —l —a —n —o —b —i —s — —d —i —s —t —a —n —c —e +1 = 4 (substitude)
M —i —n —k — —o —w —s —k —i — — —d —i —s —t —a —n —c —e

M —a —h —a —l —a —n —o —b —i —s — —d —i —s —t —a —n —c —e +1 = 5 (substitude)
M —i —n —k —o — —w —s —k —i — — —d —i —s —t —a —n —c —e

M —a —h —a —l —a —n —o —b —i —s — —d —i —s —t —a —n —c —e +1 = 6 (substitude)
M —i —n —k —o —w — —s —k —i — — —d —i —s —t —a —n —c —e

M —a —h —a —l —a —n —o —b —i —s — —d —i —s —t —a —n —c —e +1 = 7 (substitude)
M —i —n —k —o —w —s — —k —i — — —d —i —s —t —a —n —c —e

M —a —h —a —l —a —n —o —b —i —s — —d —i —s —t —a —n —c —e +1 = 8 (insert)
M —i —n —k —o —w —s —k — —i — — —d —i —s —t —a —n —c —e

M —a —h —a —l —a —n —o —b — i —s — —d —i —s —t —a —n —c —e +1 = 9 (insert)
M —i —n —k —o —w —s —k — — i — — —d —i —s —t —a —n —c —e

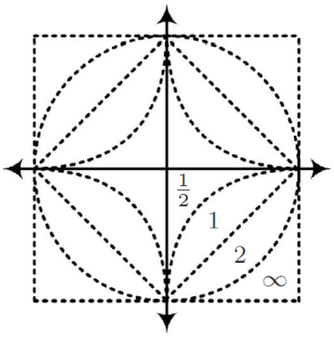


Fig. 2b: The Minkowski distances

d) Could you give three examples from the Minkowski distances group?
Not only give the name, but also the parameter instantiation!

The Minkowski distance is the generalized Lp-norm of the difference (p. 143).
Lp-norm stands for Lebesgue measure (not in book).

1) The case p= ∞ is the Chebyshev or Chessboard[1] or Supremum[2] or Maximum [3] or Moore[4] distance
2) The case p=2 is the Euclidean distance
3) The case p=1 is the city block distance or Manhattan or Von Neumann[4] distance
4) The case p=½ is used by Psychologists as a non-metric measure[4] to reflect human perception of distance (?affine perspective of Koenderink 1990?). Andras Hajdu[4] indicates that non-metrical distance functions started to receive growing interest in digital image processing, e.g. in image database retrieval (Jacobs et al., 2000). According to Jacobs non-metric distances are also used in string (DNA) matching and matching prototypical customers.

Question 3

Semantic scene understanding is a cornerstone for autonomous robot navigation in real-world environments [3]. For navigation in forested environments, robots must make complex decisions. In particular, there are obstacles that the robot can drive over, such as tall grass or bushes, but these must be distinguished safely from obstacles that the robot must avoid, such as boulders or tree trunks. In forested environments, one can exploit the presence of chlorophyll in certain obstacles as a way to discern which obstacles can be driven over. The detection of chlorophyll can be enhanced by analyzing the scene at multiple wavelength; a multispectral approach [3].

[1] https://numerics.mathdotnet.com/distance.html
[2] https://www.coursera.org/learn/cluster-analysis/lecture/CsCMY/2-2-distance-on-numeric-data-minkowski-distance
[3] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/dist.html
[4] http://www.sciencedirect.com/science/article/pii/S016786550800007X

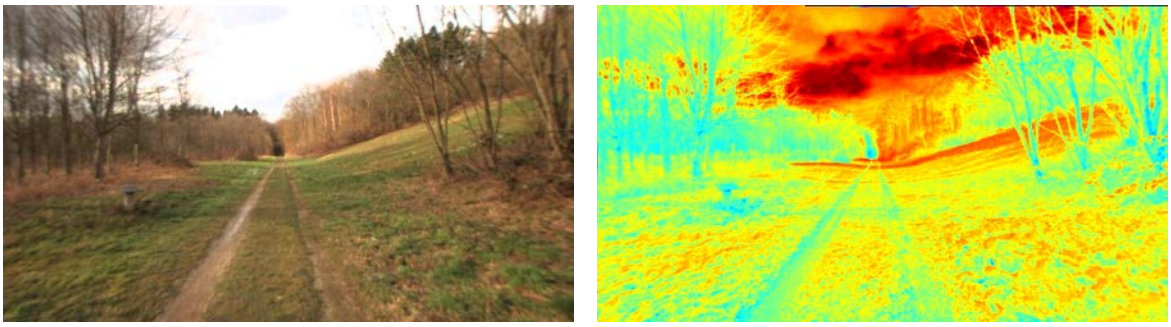


Fig 3a: A path through the forest with a color (RGB) camera and near-infrared (NIR) camera [4]

Professional near-infrared (NIR) camera systems exist, but an inexpensive alternative could be built by modifying a dashcam. From the dashcam the NIR-cut filter is removed and replaced with a Kodak Wratten 25A filter which captures the NIR in the blue and green channels.

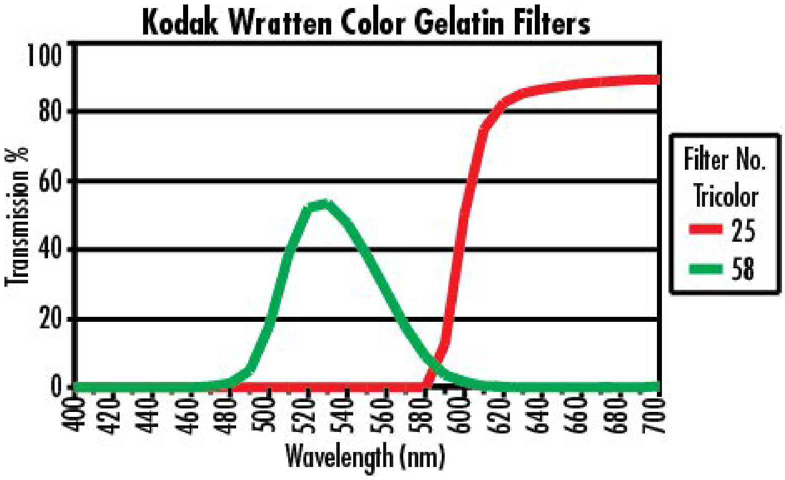


Fig 3b: The transmission of the Kodak Wratten 25 and 58 filter.

a) Kodak has not only the Wratten 25 but also the Wratten 58 filter in its collection. Could you explain why the Wratten 58 filter could be very useful if you are studying human vision?

The spectrum of the Wratten 58 filter closely resembles the cone response of the green cones of the human eye, as illustrated in Fig. 5.3 on page 79 of the book of Horst Eidenberger [1]).

The presence of vegetation in images could be even better estimated, when both the RGB and NIR color channels are combined into new color spaces as the Normalized Difference Vegetation Index (NDVI).

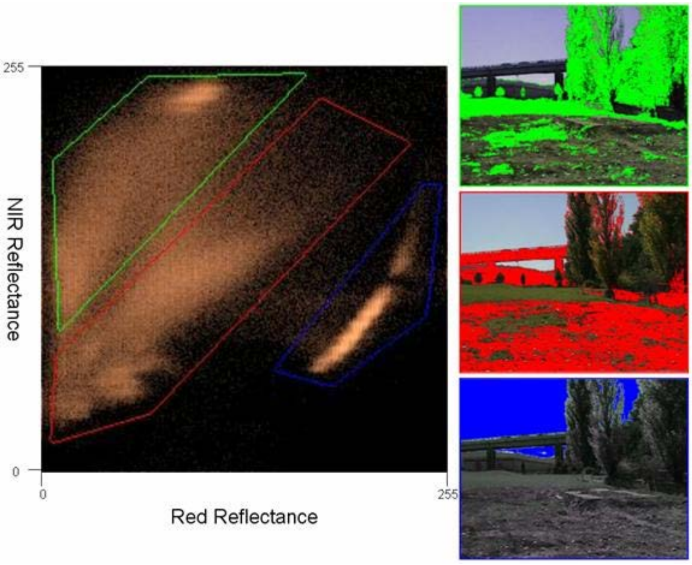


Fig 3c: Scatter plot of NIR reflectance vs. red reflectance for all pixels in a typical image [4].

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.

b) Are the green, red and blue regions in Fig. 3c an example of Hedging or Separation? **Explain your answer!**

The regions are a clear example of Hedging, as explained on page 198 of the book of Horst Eidenberger [1]). Note that the pixels could also be easily separated by two diagonal lines.

The Normalized Difference Vegetation Index (NDVI) could be calculated with the following formula:

$$\mathrm{NDVI} = \frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{RED}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{RED}}}$$

c) Explain how this formula corresponds with the pattern visible in the scatter plot (Fig 3c)?

The NDVI formula transforms the color-space along the diagonal of Fig 3c.

The multispectral images are fed into a Deep Convolutional Neural Network (DCNN), with 13 layers at the contractive side and five up-convolutional layers at the expansive side, as illustrated in Fig 3d:

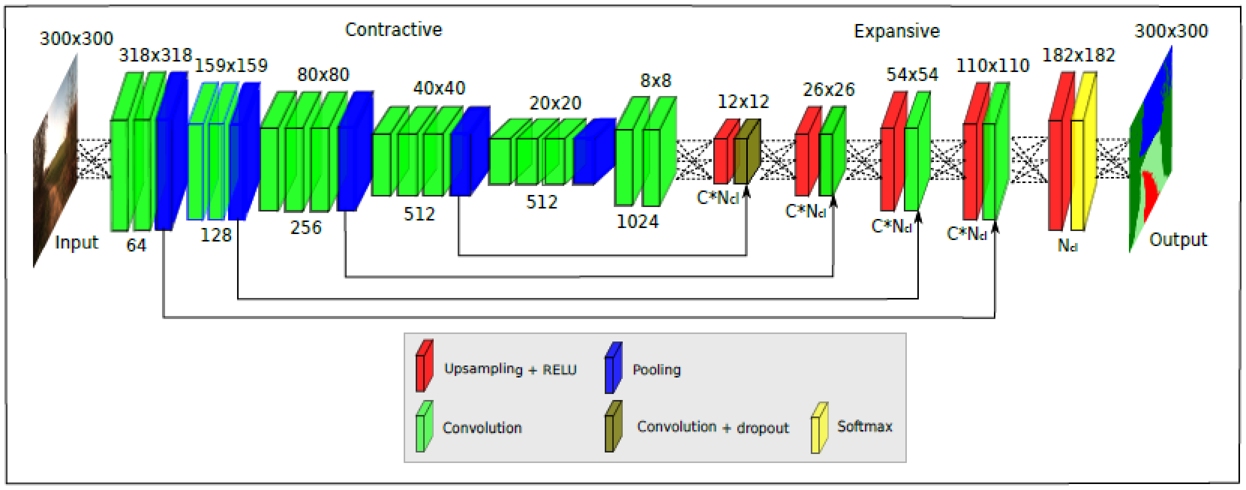


Fig 3d: The Convolutional Neural Network architecture for Semantic Scene Understanding [3].

d) Are the 13 layers at the contractive side of Fig 3d an example of spatial autocorrelation, temporal autocorrelation or crosscorrelation? **Explain your answer!**

This is a clear example of spatial autocorrelation, because the effect of surrounding pixels are taken into account. The images are analyzed, no video streams, so no temporal information is provided. The images are not correlated against templates or other media-streams (in [3] the fusion between NIR and RGB-streams is performed at the end, by combining two of the networks illustrated in Fig 3d).

The results of the semantic interpretation of the scene by this DCNN are quite impressive, as can be seen in Fig 3e. This interpretation is mainly based on multispectral colors.

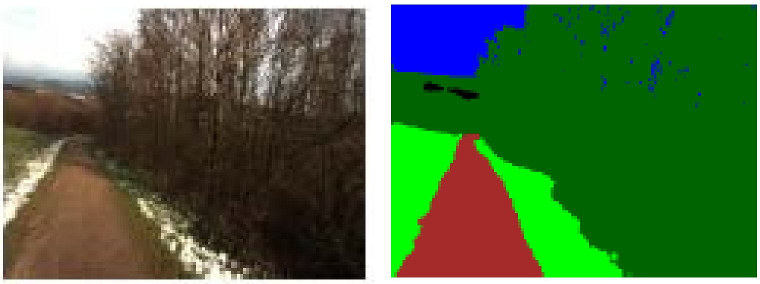


Fig 3e: The result of the semantic interpretation of a forest scene, with red indicating the trail, bright green indicating the grass and dark green indicating the vegetation [3].

e) Which two other general properties could be extracted from objects in images, other than color? **Think of classical computer vision descriptions!**

1) Texture (section 5.3 of the book; pp. 86-90) &
2) Shape and Spatial Relationships (section 5.4 of the book; pp. 90-95).

f) Could these properties be represented as features discovered in a DCNN like Fig 3d? **Explain how!**

Yes, the spatial autocorrelation in the contractive phase can be trained to find spatial patterns, which could be both texture patterns and edge / shape patterns.

Question 4

A recent article [5] describes how hearing aids can be improved to segregate sounds. Even modern hearing aids have the tendency to amplify all sounds at once, which makes it difficult to distinguish a voice from background noise, leading to the “cocktail party problem”.

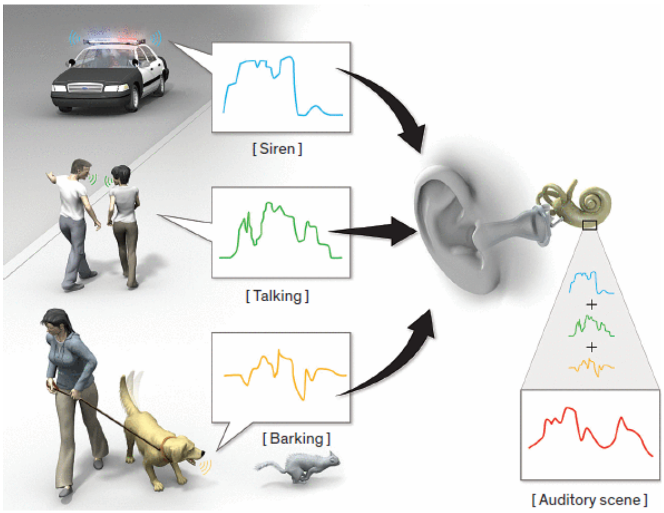


Fig 4a: The human ear can separate the sound streams coming from several sources, such as separating the Talking-stream from Barking- and Siren-stream (Courtesy DeLiang Wang [5]).

DeLiang Wang applied a deep learning approach to classify the different sound streams into speech and non-speech. This is done by analyzing the time-frequency representation of the sound.

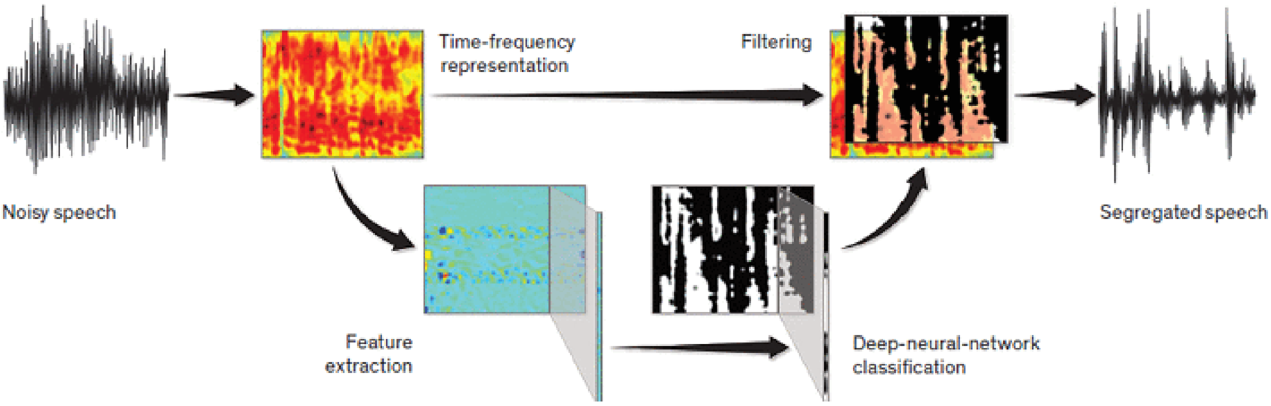


Fig 4b: To separate speech from noise, the sound is sampled in time-frequency units, followed by a feature extraction specific for speech (Courtesy DeLiang Wang [5]).

The book of Horst Eidenberger [1] describes three methods to identify the fundamental frequency of a sound. The simplest method is the Zero Crossing Rate. More advanced methods described in the book are based on *auto-regression* and *auto-correlation.* In this case estimation of the pitch is more important than estimation of the rhythm.

a) Should one use for the estimation of the fundamental frequency (calculation of the pitch histogram) an *auto-regression* or *auto-correlation* approach? **Explain your answer!**

The calculation of the fundamental frequency with a pitch histogram is an application of autocorrelation, because it uses a positive convolution (the maximum value indicates the dominant wavelet). Autoregression (i.e the Linear Predictive Coding approach) is more appropriate to detect more complex patterns in sound .

The time-frequency representation is analyzed on 85 distinctive features, such as loudness, harmonic structure and onset. Comparable with the onset is estimation of the attack and release phase of a sound in the MPEG-7 standard. Actually, the *log attack time* is estimated in the MPEG-7 standard.

b) Has taking the logarithm of the attack time the effect that differences in the sharpness of sounds are stressed or reduced? **Explain your answer!**

By taking the logarithm of the time, differences in the amplitude are stressed.

c) Measuring the attack time is maximum-based feature transformation.
Which maximum-based feature transformations are applied in image processing and stock analysis?

On page 196 several outstanding maximum-based feature transformation are listed:
* attack time
* dominant colors
* edge extraction
* resistance / support lines
The later three bullits are all good answers

The deep-learning network for the classification of speech or non-speech was trained by supervised learning: providing the network with examples how the speech should be separated from the background noise. Actually, the network was trained for five 20-ms successive frames, each with 10-ms overlap, to incorporate temporal context of changes in the background noise while a syllable was spoken.

d) Would it be beneficial to use convolutional layers inside this network?
Explain your answer!

Yes, convolutional layers make it possible to detect the temporal autocorrelation in the audio stream.

Question 5

Emotions can be measured by interpreting many different media and signals.

a) Name five different ways to measure emotions from humans.
In addition to the name, also explain in a few sentences the technique behind the measurement!

See slide 17/63 from BioSignals lecture:
* Sentiment
* Facial expressions
* Voice
* Posture
* Heart rate
* Skin conductance
* Muscle activity
* Pupil size.

An often used physiological measurement within Human-Computer Interaction is Galvanic Skin Response (also called Electrodermal Activity).

b) Explain what Galvanic Skin Response is.

See slide 28/63 from BioSignals lecture:
GSR is highly sensitive to emotions in some people. Phobia, anger, startle response, orienting response and sexual feelings are all among the emotions which may produce similar GSR responses.:

c) Explain for which mental states it is indicative.

See slide 28/63 from BioSignals lecture:
A method of measuring the electrical resistance of the skin:
- Attach two electrodes to the skin
- Acquire a base measure
- As the activity being studied is performed, recordings are made from the electrodes

d) Explain how it can be measured.

See slide 28/63 from BioSignals lecture:

- Attach two electrodes to the skin
- Acquire a base measure
- As the activity being studied is performed, recordings are made from the electrodes

Question 6

Gibbs sampling generates samples from the distribution *P(x,y)* when the conditional probabilities *P(x|y)* and *P(y|x)* are given. You can convert a conditional probability *P(x|y)* from conditional probability *P(y|x)* with Bayes rule:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

a) If you are interested in the posteriori *P(y|x),* which of the probabilities is the prior?

The prior is *P(y), P(x|y)* is the likelihood and *P(x)* is the evidence.

This process is called forward reasoning, to include new knowledge about the world when new facts (observations) arrive. When we have a large series of observations o_i and a small series of classes $c_{j\,,}$ you could build a Bayesian Classifier *BC*($\vec{o}$)= c_j based on the observation history $\vec{o} = (o_1 \wedge \cdots \wedge o_n)$ with finding the maximum conditional probability:

$$BC(\vec{o}) = \ c_j \ \textit{with } j = \arg max_c \ P(c_j|o_1 \wedge \cdots \wedge o_n)$$

A naïve Bayesian Classifier *BC*($\vec{o}$)= c_j estimates this posteriori by assuming the conditional probability can be estimated by $P(\vec{o}|c_j) = \Pi_i P(o_i|c_j)$, which means that the assumption is made that $P(o_i|c_j)$ and $P(o_{i'}|c_j)$ are independent.

b) Show for a naïve Bayesian Classifier how $P(c_j|\vec{o})$ can be derived from $P(\vec{o}|c_j)$.

See Eq (9.6) at page 161: $P(c_j|\vec{o})$.= $P(\vec{o}|c_j) \cdot \frac{P(c_j)}{P(\vec{o})}$.= $\frac{\Pi_i P(o_i|c_j)\cdot P(c_j)}{\sum_k P(\vec{o}|c_k)\cdot P(c_k)}$

You reached the end of the exam!